Preface
Introductory Statistics is intended for the one-semester introduction to statistics course for students who are not mathematics or engineering majors. It focuses on the interpretation of statistical results, especially in real world settings, and assumes that students have an understanding of intermediate algebra. In addition to end of section practice and homework sets, examples of each topic are explained step-by-step throughout the text and followed by a Try It problem that is designed as extra practice for students. This book also includes collaborative exercises and statistics labs designed to give students the opportunity to work together and explore key concepts. To support today's student in understanding technology, this book features TI 83, 83+, 84, or 84+ calculator instructions at strategic points throughout. While the book has been built so that each chapter builds on the previous, it can be rearranged to accommodate any instructor's particular needs.

Welcome to *Introductory Statistics*, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

The foundation of this textbook is *Collaborative Statistics*, by Barbara Illowsky and Susan Dean. Additional topics, examples, and innovations in terminology and practical applications have been added, all with a goal of increasing relevance and accessibility for students.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 25 books for college and AP® courses used by hundreds of thousands of students. OpenStax Tutor, our low-cost personalized learning tool, is being used in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax's resources

Customization

Introductory Statistics is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit your book page on OpenStax.org for more information.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on OpenStax.org.

Format

You can access this textbook for free in web view or PDF through OpenStax.org, and in low-cost print and iBooks editions.

About *Introductory Statistics*

Introductory Statistics follows scope and sequence requirements of a one-semester introduction to statistics course and is geared toward students majoring in fields other than math or engineering. The text assumes some knowledge of intermediate algebra and focuses on statistics application over theory. *Introductory Statistics* includes innovative practical applications that make the text relevant and accessible, as well as collaborative exercises, technology integration problems, and statistics labs.

Coverage and scope

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 Continuous Random Variables
Chapter 6 The Normal Distribution
Chapter 7 The Central Limit Theorem
Chapter 8 Confidence Intervals
Chapter 9 Hypothesis Testing with One Sample
Chapter 10 Hypothesis Testing with Two Samples
Chapter 11 The Chi-Square Distribution
Chapter 12 Linear Regression and Correlation
Chapter 13 F Distribution and One-Way ANOVA

Alternate sequencing

Introductory Statistics was conceived and written to fit a particular topical sequence, but it can be used flexibly to accommodate other course structures. One such potential structure, which fits reasonably well with the

textbook content, is provided below. Please consider, however, that the chapters were not written to be completely independent, and that the proposed alternate sequence should be carefully considered for student preparation and textual consistency.

Chapter 1 Sampling and Data
Chapter 2 Descriptive Statistics
Chapter 12 Linear Regression and Correlation
Chapter 3 Probability Topics
Chapter 4 Discrete Random Variables
Chapter 5 Continuous Random Variables
Chapter 6 The Normal Distribution
Chapter 7 The Central Limit Theorem
Chapter 8 Confidence Intervals
Chapter 9 Hypothesis Testing with One Sample
Chapter 10 Hypothesis Testing with Two Samples
Chapter 11 The Chi-Square Distribution
Chapter 13 F Distribution and One-Way ANOVA

Pedagogical foundation and features

- **Examples** are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum of practical topics, including examples about college life and learning, health and medicine, retail and business, and sports and entertainment.
- **Try It** practice problems immediately follow many examples and give students the opportunity to practice as they read the text. **They are usually based on practical and familiar topics, like the Examples themselves**.
- **Collaborative Exercises** provide an in-class scenario for students to work together to explore presented concepts.
- **Using the TI-83, 83+, 84, 84+ Calculator** shows students step-by-step instructions to input problems into their calculator.

The Technology Icon indicates where the use of a TI calculator or computer software is recommended.
Practice, Homework, and Bringing It Together problems give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.

Statistics labs

These innovative activities were developed by Barbara Illowsky and Susan Dean in order to offer students the experience of designing, implementing, and interpreting statistical analyses. They are drawn from actual experiments and data-gathering processes and offer a unique hands-on and collaborative experience. The labs provide a foundation for further learning and classroom interaction that will produce a meaningful application of statistics.

Statistics Labs appear at the end of each chapter and begin with student learning outcomes, general estimates for time on task, and any global implementation notes. Students are then provided with step-by-step guidance, including sample data tables and calculation prompts. The detailed assistance will help the students successfully apply the concepts in the text and lay the groundwork for future collaborative or individual work.

Additional resources

Student and instructor resources

We’ve compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Community Hubs

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons – a platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty.

To reach the Community Hubs, visit www.oercommons.org/hubs/OpenStax.

Partner resources

OpenStax Partners are our allies in the mission to make high-quality learning materials affordable and accessible to students and instructors everywhere. Their tools integrate seamlessly with our OpenStax titles at a low cost. To access the partner resources for your text, visit your book page on OpenStax.org.

About the authors

Senior contributing authors

Barbara Illowsky, De Anza College
Susan Dean, De Anza College

Contributing authors

Birgit Aquilonius, West Valley College
Charles Ashbacher, Upper Iowa University, Cedar Rapids

Abraham Biggs, Broward Community College
Daniel Birmajer, Nazareth College
Roberta Bloom, De Anza College
Bryan Blount, Kentucky Wesleyan College
Ernest Bonat, Portland Community College
Sarah Boslaugh, Kennesaw State University
David Bosworth, Hutchinson Community College
Sheri Boyd, Rollins College
George Bratton, University of Central Arkansas
Jing Chang, College of Saint Mary
Laurel Chiappetta, University of Pittsburgh
Lenore Desilets, De Anza College
Matthew Einsohn, Prescott College
Ann Flanigan, Kapiolani Community College
David French, Tidewater Community College
Mo Geraghty, De Anza College
Larry Green, Lake Tahoe Community College
Michael Greenwich, College of Southern Nevada
Inna Grushko, De Anza College
Valier Hauber, De Anza College
Janice Hector, De Anza College
Jim Helmreich, Marist College
Robert Henderson, Stephen F. Austin State University
Mel Jacobsen, Snow College
Mary Jo Kane, De Anza College
Lynette Kenyon, Collin County Community College
Charles Klein, De Anza College
Alexander Kolovos
Sheldon Lee, Viterbo University
Sara Lenhart, Christopher Newport University
Wendy Lightheart, Lane Community College
Vladimir Logvenenko, De Anza College
Jim Lucas, De Anza College
Lisa Markus, De Anza College
Miriam Masullo, SUNY Purchase
Diane Mathios, De Anza College
Robert McDevitt, Germanna Community College

Mark Mills, Central College
Cindy Moss, Skyline College
Nydia Nelson, St. Petersburg College
Benjamin Ngwudike, Jackson State University
Jonathan Oaks, Macomb Community College
Carol Olmstead, De Anza College
Adam Pennell, Greensboro College
Kathy Plum, De Anza College
Lisa Rosenberg, Elon University
Sudipta Roy, Kankakee Community College
Javier Rueda, De Anza College
Yvonne Sandoval, Pima Community College
Rupinder Sekhon, De Anza College
Travis Short, St. Petersburg College
Frank Snow, De Anza College
Abdulhamid Sukar, Cameron University
Jeffery Taub, Maine Maritime Academy
Mary Teegarden, San Diego Mesa College
John Thomas, College of Lake County
Philip J. Verrecchia, York College of Pennsylvania
Dennis Walsh, Middle Tennessee State University
Cheryl Wartman, University of Prince Edward Island
Carol Weideman, St. Petersburg College
Andrew Wiesner, Pennsylvania State University

Introduction
class="introduction"

We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.
- Create and interpret frequency tables.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

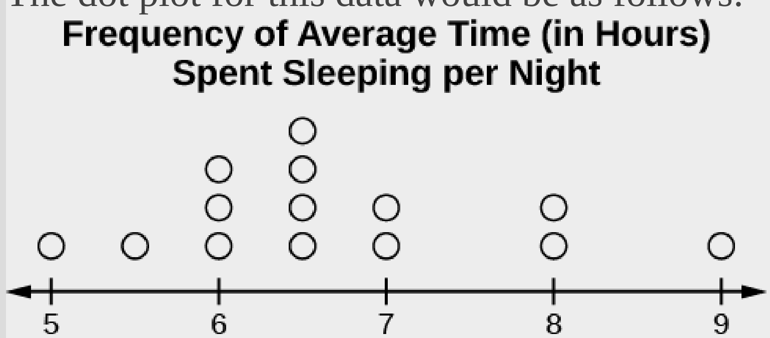
Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

Definitions of Statistics, Probability, and Key Terms

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

Note:
Collaborative Exercise
In your classroom, try this exercise. Have class members write down the average time (in hours, to the nearest half-hour) they sleep per night. Your instructor will record the data. Then create a simple graph (called a **dot plot**) of the data. A dot plot consists of a number line and dots (or points) positioned above the number line. For example, consider the following data:
5 5.5 6 6 6 6.5 6.5 6.5 6.5 7 7 8 8 9
The dot plot for this data would be as follows:



Does your dot plot look the same as or different from the example? Why? If you did the same example in an English class with the same number of students, do you think the results would be the same? Why or why not? Where do your data appear to cluster? How might you interpret the clustering?
The questions above ask you to analyze and interpret your data. With this example, you have begun your study of statistics.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example,

finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

Probability

Probability is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a **fair** coin four times, the outcomes may not be two heads and two tails. However, if you toss the same coin 4,000 times, the outcomes will be close to half heads and half tails. The expected theoretical probability of heads in any one toss is – or 0.5. Even though the outcomes of a few repetitions are uncertain, there is a regular pattern of outcomes when there are many repetitions. After reading about the English statistician Karl **Pearson** who tossed a coin 24,000 times with a result of 12,012 heads, one of the authors tossed a coin 2,000 times. The results were 996 heads. The fraction —— is equal to 0.498 which is very close to 0.5, the expected probability.

The theory of probability began with the study of games of chance such as poker. Predictions take the form of probabilities. To predict the likelihood of an earthquake, of rain, or whether you will get an A in this course, we use probabilities. Doctors use probability to determine the chance of a vaccination causing the disease the vaccination is supposed to prevent. A stockbroker uses probability to determine the rate of return on a client's investments. You might use probability to decide to buy a lottery ticket or

not. In your study of statistics, you will use the power of mathematics through probability calculations to analyze and interpret your data.

Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population. Data are the result of sampling from a population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In presidential elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

From the sample data, we can calculate a statistic. A **statistic** is a number that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the average number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter. A **parameter** is a numerical characteristic of the whole population that can be estimated by a statistic. Since we considered all math classes to be the population, then the average number of points earned per student over all the math classes is an example of a parameter.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a **representative sample**. We are interested in both the sample statistic and the population parameter in

inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

A **variable**, usually notated by capital letters such as *X* and *Y*, is a characteristic or measurement that can be determined for each member of a population. Variables may be **numerical** or **categorical**. **Numerical variables** take on values with equal units such as weight in pounds and time in hours. **Categorical variables** place the person or thing into a category. If we let *X* equal the number of points earned by one math student at the end of a term, then *X* is a numerical variable. If we let *Y* be a person's party affiliation, then some examples of *Y* include Republican, Democrat, and Independent. *Y* is a categorical variable. We could do some math with values of *X* (calculate the average number of points earned, for example), but it makes no sense to do math with values of *Y* (calculating an average party affiliation makes no sense).

Data are the actual values of the variable. They may be numbers or they may be words. **Datum** is a single value.

Two words that come up often in statistics are **mean** and **proportion**. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men and 18 are women, then the proportion of men students is — and the proportion of women students is —. Mean and proportion are discussed in more detail in later chapters.

Note:
NOTE
The words "**mean**" and "**average**" are often used interchangeably. The substitution of one word for the other is common practice. The technical term is "arithmetic mean," and "average" is technically a center location. However, in practice among non-statisticians, "average" is commonly accepted for "arithmetic mean."

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Solution:

The **population** is all first year students attending ABC College this term.

The **sample** could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample may not represent the entire population).

The **parameter** is the average (mean) amount of money spent (excluding books) by first year college students at ABC College this term.

The **statistic** is the average (mean) amount of money spent (excluding books) by first year college students in the sample.

The **variable** could be the amount of money spent (excluding books) by one first year student. Let X = the amount of money spent (excluding books) by one first year student attending ABC College.

The **data** are the dollar amounts spent by the first year students. Examples of the data are $150, $200, and $225.

Note:
Try It

Exercise:

Problem:

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent \$65, \$75, and \$95, respectively.

Solution:
Try It Solutions

The **population** is all families with children attending Knoll Academy.

The **sample** is a random selection of 100 families with children attending Knoll Academy.

The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.

The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.

The **data** are the dollar amounts spent by the families. Examples of the data are \$65, \$75, and \$95.

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. Population_____ 2. Statistic _____ 3. Parameter _____ 4. Sample _____ 5. Variable _____ 6. Data _____

- a) all students who attended the college last year
- b) the cumulative GPA of one student who graduated from the college last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the college last year, randomly selected
- e) the average cumulative GPA of students who graduated from the college last year
- f) all students who graduated from the college last year
- g) the average cumulative GPA of students in the study who graduated from the college last year

Solution:

1. f 2. g 3. e 4. d 5. b 6. c

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

Solution:

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** X = the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Example:
Exercise:

Problem:

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Solution:

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** X = the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

Note:
Collaborative Exercise

Do the following exercise collaboratively with up to four people per group. Find a population, a sample, the parameter, the statistic, a variable, and data for the following study: You want to determine the average (mean) number of glasses of milk college students drink per day. Suppose yesterday, in your English class, you asked five students how many glasses of milk they drank the day before. The answers were 1, 0, 1, 3, and 4 glasses of milk.

References

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

Chapter Review

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

Practice

Use the following information to answer the next five exercises. Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once they start the treatment. Two researchers each follow a different set of 40 patients with AIDS from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A:
3 4 11 15 16 17 22 44 37 16 14 24 25 15 26 27 33 29 35 44 13 21 22 10 12 8 40 32 26 27 31 34 29 17 8 24 18 47 33 34

Researcher B:
3 14 11 5 16 17 28 41 31 18 14 14 26 25 21 22 31 2 35 44 23 21 21 16 12 18 41 22 16 25 33 34 29 13 18 24 23 42 33 29

Determine what the key terms refer to in the example for Researcher A.

Exercise:

Problem: population

Solution:

AIDS patients.

Exercise:

Problem: sample

Exercise:

Problem: parameter

Solution:

The average length of time (in months) AIDS patients live after treatment.

Exercise:

Problem: statistic

Exercise:

Problem: variable

Solution:

X = the length of time (in months) AIDS patients live after treatment

HOMEWORK

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

Exercise:

Problem:

A fitness center is interested in the mean amount of time a client exercises in the center each week.

Exercise:

Problem:

Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

Solution:

a. all children who take ski or snowboard lessons
b. a group of these children
c. the population mean age of children who take their first snowboard lesson
d. the sample mean age of children who take their first snowboard lesson
e. X = the age of one child who takes his or her first ski or snowboard lesson
f. values for X, such as 3, 7, and so on

Exercise:

Problem:

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Exercise:

Problem:

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

Solution:

a. the clients of the insurance companies
b. a group of the clients
c. the mean health costs of the clients
d. the mean health costs of the sample
e. X = the health costs of one client
f. values for X, such as 34, 9, 82, and so on

Exercise:

Problem:

A politician is interested in the proportion of voters in his district who think he is doing a good job.

Exercise:

Problem:

A marriage counselor is interested in the proportion of clients she counsels who stay married.

Solution:

a. all the clients of this counselor
b. a group of clients of this marriage counselor
c. the proportion of all her clients who stay married
d. the proportion of the sample of the counselor’s clients who stay married
e. X = the number of couples who stay married
f. yes, no

Exercise:

Problem:

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

Exercise:

Problem:

A marketing company is interested in the proportion of people who will buy a particular product.

Solution:

a. all people (maybe in a certain geographic area, such as the United States)
b. a group of the people
c. the proportion of all people who will buy the product
d. the proportion of the sample who will buy the product
e. X = the number of people who will buy it
f. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

Exercise:

Problem: What is the population she is interested in?

a. all Lake Tahoe Community College students
b. all Lake Tahoe Community College English students
c. all Lake Tahoe Community College students in her classes
d. all Lake Tahoe Community College math students

Exercise:

Problem: Consider the following:

= number of days a Lake Tahoe Community College math student is absent

In this case, *X* is an example of a:

a. variable.
b. population.
c. statistic.
d. data.

Solution:

a

Exercise:

Problem:

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

a. parameter.
b. data.
c. statistic.
d. variable.

Glossary

Average
: also called mean; a number that describes the central tendency of the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data can be put into two groups: **qualitative** (an attribute whose value is indicated by a label) or **quantitative** (an attribute whose value is indicated by a number). Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population

all individuals, objects, or measurements whose properties are being studied

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion

the number of successes divided by the total number in the sample

Representative Sample

a subset of the population that has the same characteristics as the population

Sample

a subset of the population studied

Statistic

a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable

a characteristic of interest for each person or object in a population

Data, Sampling, and Variation in Data and Sampling

Data may come from a population or from a sample. Lowercase letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

Qualitative data are the result of categorizing or describing attributes of a population. Qualitative data are also often called **categorical data**. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair color or blood type.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the lengths in minutes for all the phone calls that you make in a week, with numbers like 2.4, 7.5, or 11.0, would be quantitative continuous data.

Example:
Data Sample of Quantitative Discrete Data
The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Note:
Try It
Exercise:

Problem:

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Solution:
Try It Solutions

quantitative discrete data

Example:
Data Sample of Quantitative Continuous Data
The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data.

Note:
Try It
Exercise:

Problem:

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Solution:
Try It Solutions

quantitative continuous data

Example:
You go to the supermarket and purchase three cans of soup (19 ounces tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces pistachio ice cream and 32 ounces chocolate chip cookies).
Exercise:

Problem:

Name data sets that are quantitative discrete, quantitative continuous, and qualitative.

Solution:

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are qualitative data because they are categorical.

Try to identify additional data sets in this example.

Example:
The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are qualitative data.

Note:
Try It
Exercise:

Problem:

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Solution:
Try It Solutions

qualitative data

Note:
Note
You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F.

Example:
Exercise:

Problem:

Work collaboratively to determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words "the number of."

a. the number of pairs of shoes you own
b. the type of car you drive
c. the distance it is from your home to the nearest grocery store
d. the number of classes you take per school year.
e. the type of calculator you use
f. weights of sumo wrestlers
g. number of correct answers on a quiz
h. IQ scores (This may cause some discussion.)

Solution:

Items a, d, and g are quantitative discrete; items c, f, and h are quantitative continuous; items b and e are qualitative, or categorical.

Note:
Try It
Exercise:

Problem:

Determine the correct data type (quantitative or qualitative) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

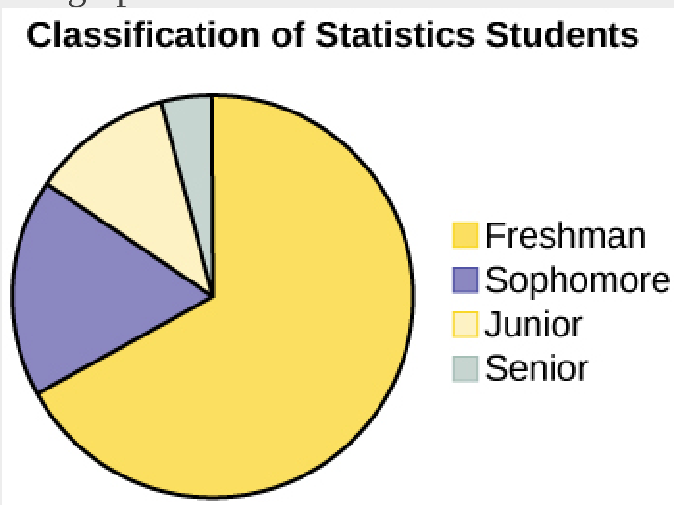
Solution:
Try It Solutions

quantitative discrete

Example:
Exercise:

Problem:

A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart [link]. What type of data does this graph show?



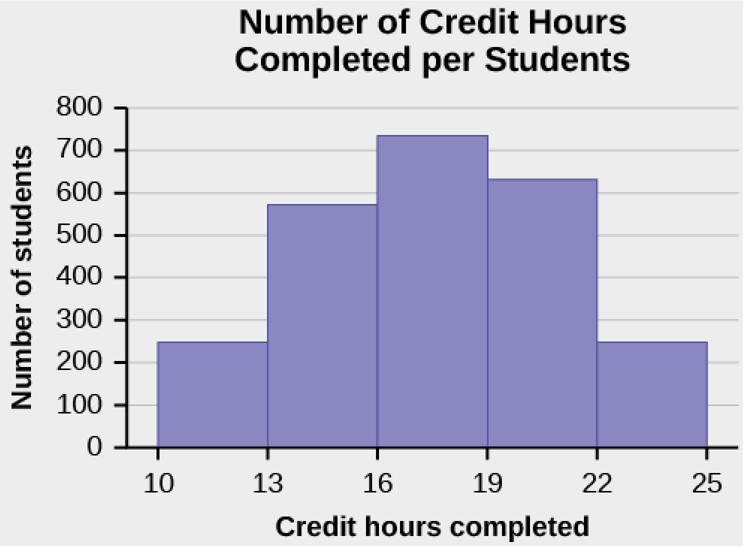
Solution:

This pie chart shows the students in each year, which is **qualitative (or categorical) data**.

Note:
Try It
Exercise:

Problem:

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.



What type of data does this graph show?

Solution:
Try It Solutions

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

Qualitative Data Discussion

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College				Foothill College		
	Number	Percent			Number	Percent
Full-time	9,200	40.9%		Full-time	4,059	28.6%
Part-time	13,296	59.1%		Part-time	10,124	71.4%
Total	22,496	100%		Total	14,183	100%

Fall Term 2007 (Census day)

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display qualitative data are pie charts and bar graphs.

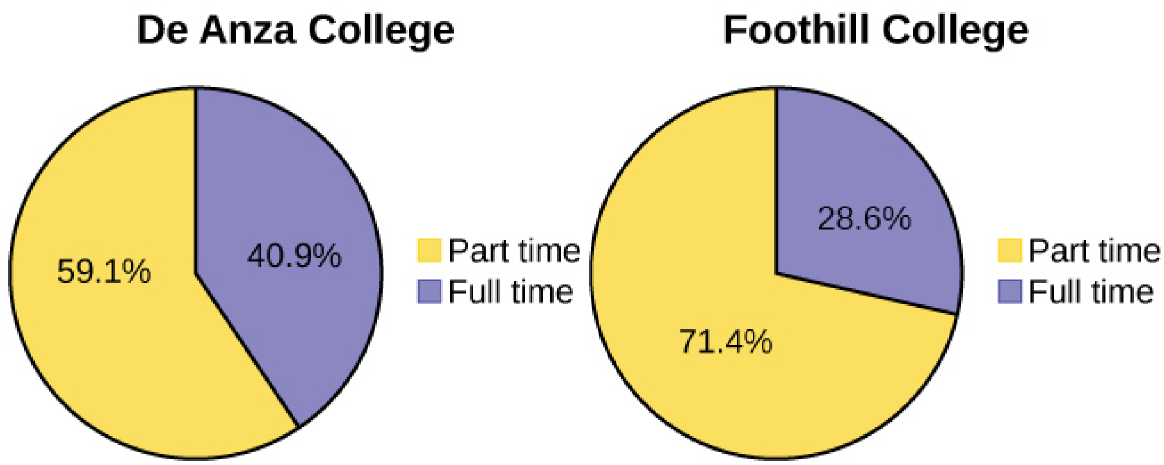
In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

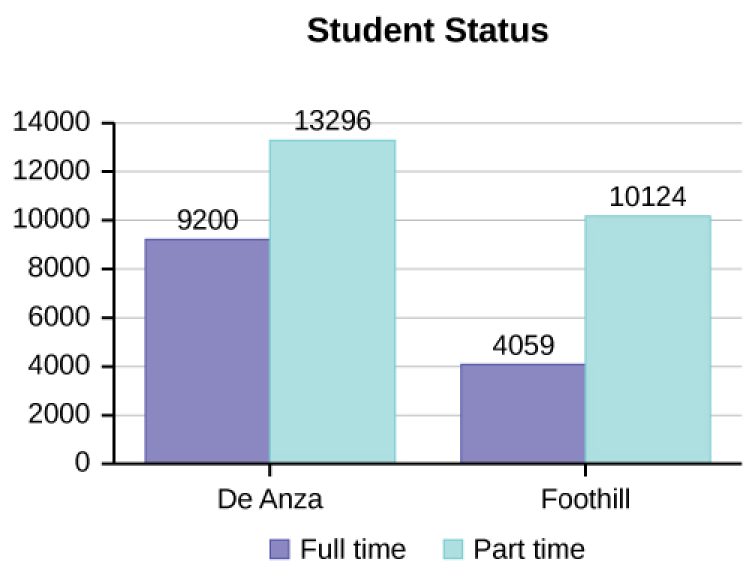
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

A **Pareto chart** consists of bars that are sorted into order by category size (largest to smallest).

Look at [link] and [link] and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the "best" graph depending on the data and the context. Our choice also depends on what we are using the data for.



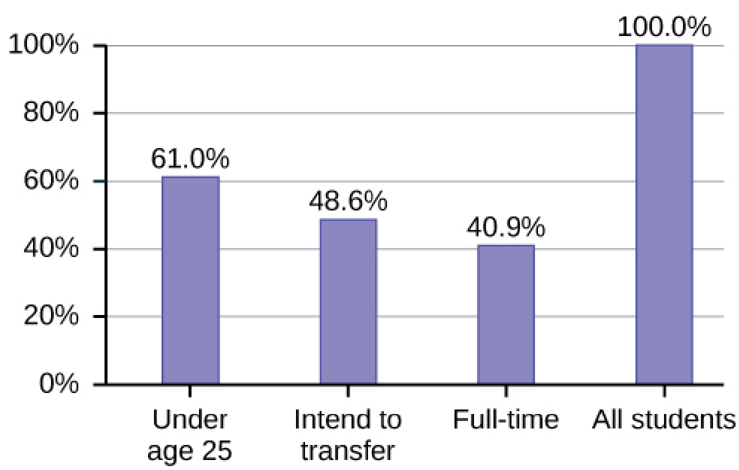


Percentages That Add to More (or Less) Than 100%

Sometimes percentages add up to be more than 100% (or less than 100%). In the graph, the percentages add to more than 100% because students can be in more than one category. A bar graph is appropriate to compare the relative size of the categories. A pie chart cannot be used. It also could not be used if the percentages added to less than 100%.

Characteristic/Category	Percent
Full-Time Students	40.9%
Students who intend to transfer to a 4-year educational institution	48.6%
Students under age 25	61.0%
TOTAL	150.5%

De Anza College Spring 2010

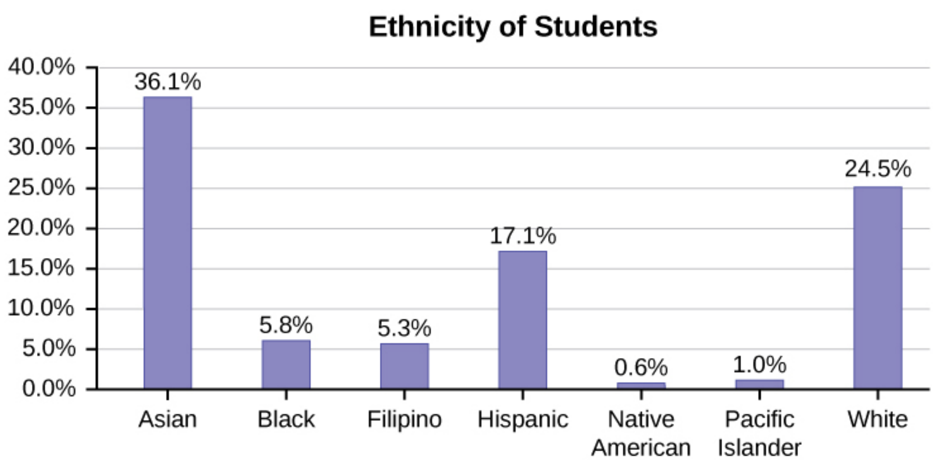


Omitting Categories/Missing Data

The table displays Ethnicity of Students but is missing the "Other/Unknown" category. This category contains people who did not feel they fit into any of the ethnicity categories or declined to respond. Notice that the frequencies do not add up to the total number of students. In this situation, create a bar graph and not a pie chart.

	Frequency	Percent
Asian	8,794	36.1%
Black	1,412	5.8%
Filipino	1,298	5.3%
Hispanic	4,180	17.1%
Native American	146	0.6%
Pacific Islander	236	1.0%
White	5,978	24.5%
TOTAL	22,044 out of 24,382	90.4% out of 100%

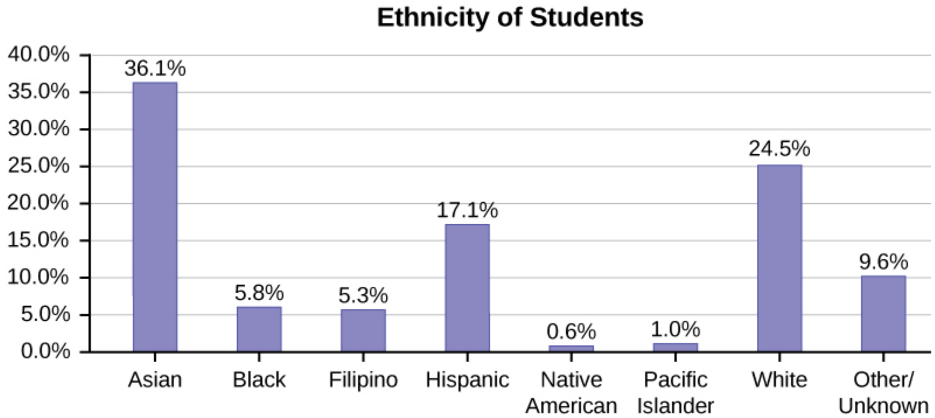
Ethnicity of Students at De Anza College Fall Term 2007 (Census Day)



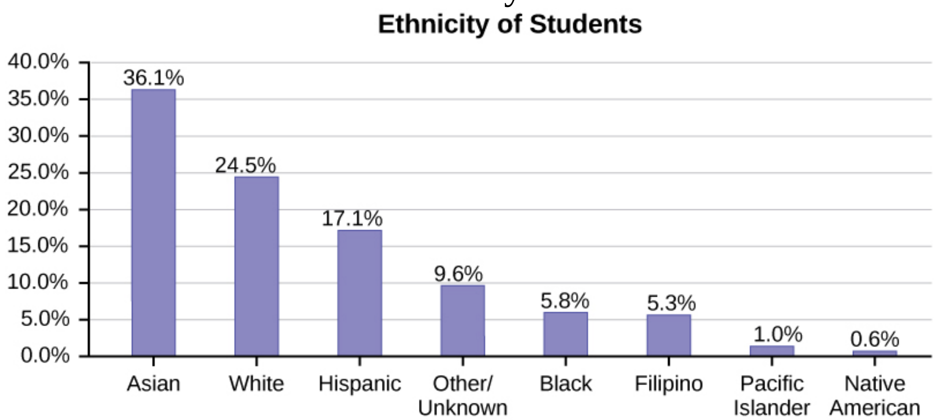
The following graph is the same as the previous graph but the "Other/Unknown" percent (9.6%) has been included. The "Other/Unknown" category is large compared to some of the other categories (Native American, 0.6%, Pacific Islander 1.0%). This is important to know when we think about what the data are telling us.

This particular bar graph in [link] can be difficult to understand visually. The graph in [link] is a Pareto chart. The Pareto chart has the bars sorted from largest to smallest and is easier to read and interpret.

Bar Graph with Other/Unknown Category

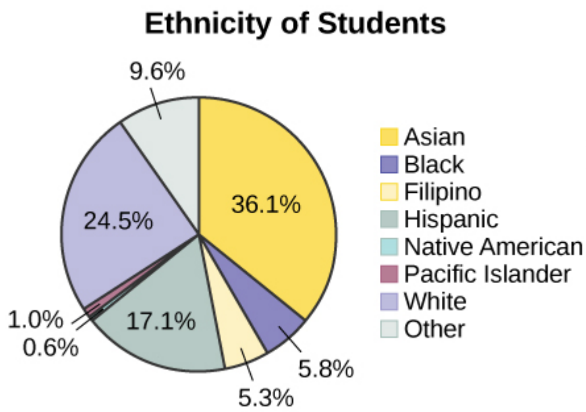


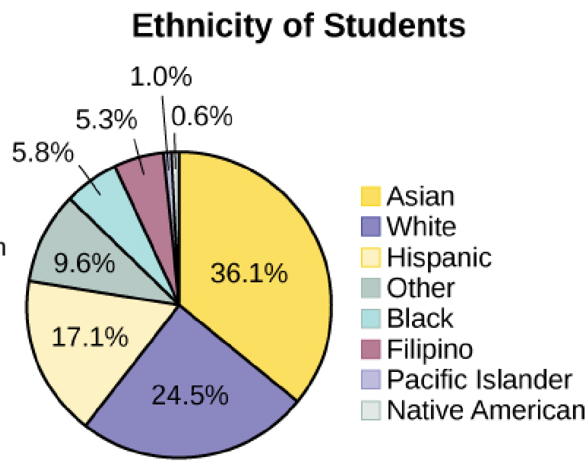
Pareto Chart With Bars Sorted by Size



Pie Charts: No Missing Data

The following pie charts have the “Other/Unknown” category included (since the percentages must add to 100%). The chart in [link] is organized by the size of each wedge, which makes it a more visually informative graph than the unsorted, alphabetical graph in [link].





Sampling

Gathering information about an entire population often costs too much or is virtually impossible. Instead, we use a sample of the population. **A sample should have the same characteristics as the population it is representing.** Most statisticians use various methods of random sampling in an attempt to achieve this goal. This section will describe a few of the most common methods. There are several different methods of **random sampling**. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. The easiest method to describe is called a **simple random sample**. Any group of *n* individuals is equally likely to be chosen as any other group of *n* individuals if the simple random sampling technique is used. In other words, each sample of the same size has an equal chance of being selected. For example, suppose Lisa wants to form a four-person study group (herself and three other people) from her pre-calculus class, which has 31 members not including Lisa. To choose a simple random sample of size three from the other members of her class, Lisa could put all 31 names in a hat, shake the hat, close her eyes, and pick out three names. A more technological way is for Lisa to first list the last names of the members of her class together with a two-digit number, as in [link]:

ID	Name	ID	Name	ID	Name
00	Anselmo	11	King	21	Roquero
01	Bautista	12	Legeny	22	Roth
02	Bayani	13	Lundquist	23	Rowell
03	Cheng	14	Macierz	24	Salangsang
04	Cuarismo	15	Motogawa	25	Slade
05	Cuningham	16	Okimoto	26	Stratcher
06	Fontecha	17	Patel	27	Tallai
07	Hong	18	Price	28	Tran
08	Hoobler	19	Quizon	29	Wai
09	Jiao	20	Reyes	30	Wood
10	Khan				

Class Roster

Lisa can use a table of random numbers (found in many statistics books and mathematical handbooks), a calculator, or a computer to generate random numbers. For this example, suppose Lisa chooses to generate random numbers from a calculator. The numbers generated are as follows:

0.94360 0.99832 0.14669 0.51470 0.40581 0.73381 0.04399

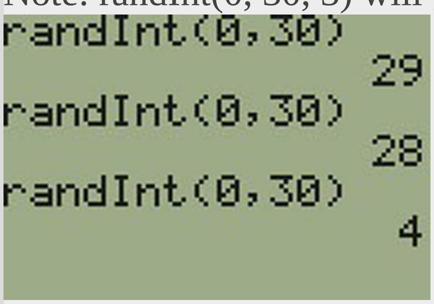
Lisa reads two-digit groups until she has chosen three class members (that is, she reads 0.94360 as the groups 94, 43, 36, 60). Each random number may only contribute one class member. If she needed to, Lisa could have generated more random numbers.

The random numbers 0.94360 and 0.99832 do not contain appropriate two digit numbers. However the third random number, 0.14669, contains 14 (the fourth random number also contains 14), the fifth random number contains 05, and the seventh random number contains 04. The two-digit number 14 corresponds to Macierz, 05 corresponds to Cuningham, and 04 corresponds to Cuarismo. Besides herself, Lisa’s group will consist of Marcierz, Cuningham, and Cuarismo.

Note:
To generate random numbers:

- Press MATH.
- Arrow over to PRB.
- Press 5:randInt(. Enter 0, 30).
- Press ENTER for the first random number.
- Press ENTER two more times for the other 2 random numbers. If there is a repeat press ENTER again.

Note: randInt(0, 30, 3) will generate 3 random numbers.



Besides simple random sampling, there are other forms of sampling that involve a chance process for getting the sample. **Other well-known random sampling methods are the stratified sample, the cluster sample, and the systematic sample.**

To choose a **stratified sample**, divide the population into groups called strata and then take a **proportionate** number from each stratum. For example, you could stratify (group) your college population by department and then choose a proportionate simple random sample from each stratum (each department) to

get a stratified random sample. To choose a simple random sample from each department, number each member of the first department, number each member of the second department, and do the same for the remaining departments. Then use simple random sampling to choose proportionate numbers from the first department and do the same for each of the remaining departments. Those numbers picked from the first department, picked from the second department, and so on represent the members who make up the stratified sample.

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. All the members from these clusters are in the cluster sample. For example, if you randomly sample four departments from your college population, the four departments make up the cluster sample. Divide your college faculty by department. The departments are the clusters. Number each department, and then choose four different numbers using simple random sampling. All members of the four departments with those numbers are the cluster sample.

To choose a **systematic sample**, randomly select a starting point and take every n^{th} piece of data from a listing of the population. For example, suppose you have to do a phone survey. Your phone book contains 20,000 residence listings. You must choose 400 names for the sample. Number the population 1–20,000 and then use a simple random sample to pick a number that represents the first name in the sample. Then choose every fiftieth name thereafter until you have a total of 400 names (you might have to go back to the beginning of your phone list). Systematic sampling is frequently chosen because it is a simple method.

A type of sampling that is non-random is convenience sampling. **Convenience sampling** involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others.

Sampling data should be done very carefully. Collecting data carelessly can have devastating results. Surveys mailed to households and then returned may be very biased (they may favor a certain group). It is better for the person conducting the survey to select the sample respondents.

True random sampling is done **with replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However for practical reasons, in most populations, simple random sampling is done **without replacement**. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

In a college population of 10,000 people, suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

When you analyze data, it is important to be aware of **sampling errors** and nonsampling errors. The actual process of sampling causes sampling errors. For example, the sample may not be large enough. Factors not related to the sampling process cause **nonsampling errors**. A defective counting device can cause a nonsampling error.

In reality, a sample will never be exactly representative of the population so there will always be some sampling error. As a rule, the larger the sample, the smaller the sampling error.

In statistics, **a sampling bias** is created when a sample is collected from a population and some members of the population are not as likely to be chosen as others (remember, each member of the population should have an equally likely chance of being chosen). When a sampling bias happens, there can be incorrect conclusions drawn about the population that is being studied.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. Common problems to be aware of include

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples

is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions

- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.
- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

Note:
Collaborative Exercise
As a class, determine whether or not the following samples are representative. If they are not, discuss the reasons.

1. To find the average GPA of all students in a university, use all honor students at the university as the sample.
2. To find out the most popular cereal among young people under the age of ten, stand outside a large supermarket for three hours and speak to every twentieth child under age ten who enters the supermarket.
3. To find the average annual income of all adults in the United States, sample U.S. congressmen. Create a cluster sample by considering each state as a stratum (group). By using simple random sampling, select states to be part of the cluster. Then survey every U.S. congressman in the cluster.

4. To determine the proportion of people taking public transportation to work, survey 20 people in New York City. Conduct the survey by sitting in Central Park on a bench and interviewing every person who sits next to you.
5. To determine the average cost of a two-day stay in a hospital in Massachusetts, survey 100 hospitals across the state using simple random sampling.

Example:
Exercise:

Problem:

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

a. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
b. A random number generator is used to select a student from the alphabetical listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
c. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
d. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
e. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he

encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

Solution:

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Note:
Try It
You are going to use the random number generator to generate different types of samples from the data.
This table displays six sets of quiz scores (each quiz counts 10 points) for an elementary statistics class.

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8

#1	#2	#3	#4	#5	#6
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Instructions: Use the Random Number Generator to pick samples.

Exercise:

Problem:

1. Create a stratified sample by column. Pick three quiz scores randomly from each column.
    - Number each row one through ten.
    - On your calculator, press Math and arrow over to PRB.
    - For column 1, Press 5:randInt(and enter 1,10). Press ENTER. Record the number. Press ENTER 2 more times (even the repeats). Record these numbers. Record the three quiz scores in column one that correspond to these three numbers.
    - Repeat for columns two through six.
    - These 18 quiz scores are a stratified sample.
2. Create a cluster sample by picking two of the columns. Use the column numbers: one through six.
    - Press MATH and arrow over to PRB.
    - Press 5:randInt(and enter 1,6). Press ENTER. Record the number. Press ENTER and record that number.
    - The two numbers are for two of the columns.
    - The quiz scores (20 of them) in these 2 columns are the cluster sample.
3. Create a simple random sample of 15 quiz scores.
    - Use the numbering one through 60.

- Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
- Press ENTER 15 times and record the numbers.
- Record the quiz scores that correspond to these numbers.
- These 15 quiz scores are the systematic sample.

4. Create a systematic sample of 12 quiz scores.

   - Use the numbering one through 60.
   - Press MATH. Arrow over to PRB. Press 5:randInt(and enter 1, 60).
   - Press ENTER. Record the number and the first quiz score. From that number, count ten quiz scores and record that quiz score. Keep counting ten quiz scores and recording the quiz score until you have a sample of 12 quiz scores. You may wrap around (go back to the beginning).

Example:
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

a. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
b. A pollster interviews all human resource personnel in five different high tech companies.
c. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
d. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

e. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

Solution:

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

Note:
Try It
Exercise:

Problem:

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Solution:

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Example:

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.

First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:

$128 $87 $173 $116 $130 $204 $147 $189 $93 $153

The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:

$50 $40 $36 $15 $50 $100 $40 $53 $22 $22

It is unlikely that any student is in both samples.

Exercise:

Problem:

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

Solution:

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

Exercise:

Problem:

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

Solution:

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:
$180 $50 $150 $85 $260 $75 $180 $200 $200 $150
Exercise:

Problem: c. Is the sample biased?

Solution:

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Note:
Try It
Exercise:

Problem:

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station's music concert events. 24 people said they'd prefer more talk shows, and 176 people said they'd prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Solution:
Try It Solutions

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn't represent the entire fan base and is probably biased towards people who would prefer music.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests

to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen’s and Jung’s samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are

considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Note:
Collaborative Exercise
Divide into groups of two, three, or four. Your instructor will give each group one six-sided die. Try this experiment twice. Roll one fair die (six-sided) 20 times. Record the number of ones, twos, threes, fours, fives, and sixes you get in [link] and [link] ("frequency" is the number of times a particular face of the die occurs):

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

First Experiment (20 rolls)

Face on Die	Frequency
1	
2	
3	
4	
5	
6	

Second Experiment (20 rolls)

Did the two experiments have the same results? Probably not. If you did the experiment a third time, do you expect the results to be identical to the first or second experiment? Why or why not?
Which experiment had the correct results? They both did. The job of the statistician is to see through the variability and draw appropriate conclusions.

References

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

Dominic Lusinchi, "'President' Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?" Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

"The Literary Digest Poll," Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).

"Gallup Presidential Election Trial-Heat Trends, 1936–2008," Gallup Politics http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in 2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

Chapter Review

Data are individual items of information that come from a population or sample. Data may be classified as qualitative(categorical), quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

Practice

Exercise:

Problem: “Number of times per week” is what type of data?

a. qualitative(categorical) b. quantitative discrete c. quantitative continuous

Use the following information to answer the next four exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Antonio, Texas. The first house in the neighborhood around the park was selected randomly, and then the resident of every eighth house in the neighborhood around the park was interviewed.

Exercise:

Problem: The sampling method was

a. simple random b. systematic c. stratified d. cluster

Solution:

b

Exercise:

Problem: “Duration (amount of time)” is what type of data?

a. qualitative(categorical) b. quantitative discrete c. quantitative continuous

Exercise:

Problem:

The colors of the houses around the park are what kind of data?

a. qualitative(categorical) b. quantitative discrete c. quantitative continuous

---

Solution:

a

Exercise:

Problem: The population is _______________________

Exercise:

Problem:

[link] contains the total number of deaths worldwide as a result of earthquakes from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802

Year	Total Number of Deaths
2005	88,003
2006	6,605
2007	712
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	**823,856**

Use [link] to answer the following questions.

a. What is the proportion of deaths between 2007 and 2012?
b. What percent of deaths occurred before 2001?
c. What is the percent of deaths that occurred in 2003 or after 2010?
d. What is the fraction of deaths that happened before 2012?
e. What kind of data is the number of deaths?
f. Earthquakes are quantified according to the amount of energy they produce (examples are 2.1, 5.0, 6.7). What type of data is that?
g. What contributed to the large number of deaths in 2010? In 2004? Explain.

Solution:

a. 0.5242
b. 0.03%
c. 6.86%

d. $\frac{823,088}{823,856}$
e. quantitative discrete
f. quantitative continuous
g. In both years, underwater earthquakes produced massive tsunamis.

For the following four exercises, determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

Exercise:

Problem:

A group of test subjects is divided into twelve groups; then four of the groups are chosen at random.

Exercise:

Problem:

A market researcher polls every tenth person who walks into a store.

Solution:

systematic

Exercise:

Problem:

The first 50 people who walk into a sporting event are polled on their television preferences.

Exercise:

Problem:

A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen.

Solution:

simple random

Use the following information to answer the next seven exercises: Studies are often done by pharmaceutical companies to determine the effectiveness of a treatment program. Suppose that a new AIDS antibody drug is currently under study. It is given to patients once the AIDS symptoms have revealed themselves. Of interest is the average (mean) length of time in months patients live once starting the treatment. Two researchers each follow a different set of 40 AIDS patients from the start of treatment until their deaths. The following data (in months) are collected.

Researcher A: 3; 4; 11; 15; 16; 17; 22; 44; 37; 16; 14; 24; 25; 15; 26; 27; 33; 29; 35; 44; 13; 21; 22; 10; 12; 8; 40; 32; 26; 27; 31; 34; 29; 17; 8; 24; 18; 47; 33; 34

Researcher B: 3; 14; 11; 5; 16; 17; 28; 41; 31; 18; 14; 14; 26; 25; 21; 22; 31; 2; 35; 44; 23; 21; 21; 16; 12; 18; 41; 22; 16; 25; 33; 34; 29; 13; 18; 24; 23; 42; 33; 29

Exercise:

Problem: Complete the tables using the data provided:

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
24.5–30.5			
30.5–36.5			
36.5–42.5			
42.5–48.5			

Researcher A

Survival Length (in months)	Frequency	Relative Frequency	Cumulative Relative Frequency
0.5–6.5			
6.5–12.5			
12.5–18.5			
18.5–24.5			
24.5–30.5			
30.5–36.5			
36.5-45.5			

Researcher B

Exercise:

Problem:

Determine what the key term data refers to in the above example for Researcher A.

Solution:

values for *X*, such as 3, 4, 11, and so on

Exercise:

Problem: List two reasons why the data may differ.

Exercise:

Problem:

Can you tell if one researcher is correct and the other one is incorrect? Why?

Solution:

No, we do not have enough information to make such a claim.

Exercise:

Problem: Would you expect the data to be identical? Why or why not?

Exercise:

Problem:

Suggest at least two methods the researchers might use to gather random data.

Solution:

Take a simple random sample from each group. One way is by assigning a number to each patient and using a random number generator to randomly select patients.

Exercise:

Problem:

Suppose that the first researcher conducted his survey by randomly choosing one state in the nation and then randomly picking 40 patients from that state. What sampling method would that researcher have used?

Exercise:

Problem:

Suppose that the second researcher conducted his survey by choosing 40 patients he knew. What sampling method would that researcher have used? What concerns would you have about this data set, based upon the data collection method?

Solution:

This would be convenience sampling and is not random.

Use the following data to answer the next five exercises: Two researchers are gathering data on hours of video games played by school-aged children and young adults. They each randomly sample different groups of 150 students from the same school. They collect the following data.

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	26	0.17	0.17
2–4	30	0.20	0.37
4–6	49	0.33	0.70

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
6–8	25	0.17	0.87
8–10	12	0.08	0.95
10–12	8	0.05	1

Researcher A

Hours Played per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0–2	48	0.32	0.32
2–4	51	0.34	0.66
4–6	24	0.16	0.82
6–8	12	0.08	0.90
8–10	11	0.07	0.97
10–12	4	0.03	1

Researcher B

Exercise:

Problem: Give a reason why the data may differ.

Exercise:

Problem:

Would the sample size be large enough if the population is the students in the school?

Solution:

Yes, the sample size of 150 would be large enough to reflect a population of one school.

Exercise:

Problem:

Would the sample size be large enough if the population is school-aged children and young adults in the United States?

Exercise:

Problem:

Researcher A concludes that most students play video games between four and six hours each week. Researcher B concludes that most students play video games between two and four hours each week. Who is correct?

Solution:

Even though the specific data support each researcher’s conclusions, the different results suggest that more data need to be collected before the researchers can reach a conclusion.

Exercise:

Problem:

As part of a way to reward students for participating in the survey, the researchers gave each student a gift card to a video game store. Would this affect the data if students knew about the award before the study?

Use the following data to answer the next five exercises: A pair of studies was performed to measure the effectiveness of a new software program designed to help stroke patients regain their problem-solving skills. Patients were asked to use the software program twice a day, once in the morning and once in the evening. The studies observed 200 stroke patients recovering over a period of several weeks. The first study collected the data in [link]. The second study collected the data in [link].

Group	Showed improvement	No improvement	Deterioration
Used program	142	43	15
Did not use program	72	110	18

Group	Showed improvement	No improvement	Deterioration
Used program	105	74	19
Did not use program	89	99	12

Exercise:

Problem: Given what you know, which study is correct?

Solution:

There is not enough information given to judge if either one is correct or incorrect.

Exercise:

Problem:

The first study was performed by the company that designed the software program. The second study was performed by the American Medical Association. Which study is more reliable?

Exercise:

Problem:

Both groups that performed the study concluded that the software works. Is this accurate?

Solution:

The software program seems to work because the second study shows that more patients improve while using the software than not. Even though the difference is not as large as that in the first study, the results from the second study are likely more reliable and still show improvement.

Exercise:

Problem:

The company takes the two studies as proof that their software causes mental improvement in stroke patients. Is this a fair statement?

Exercise:

Problem:

Patients who used the software were also a part of an exercise program whereas patients who did not use the software were not. Does this change the validity of the conclusions from [link]?

Solution:

Yes, because we cannot tell if the improvement was due to the software or the exercise; the data is confounded, and a reliable conclusion cannot be drawn. New studies should be performed.

Exercise:

Problem:

Is a sample size of 1,000 a reliable measure for a population of 5,000?

Exercise:

Problem:

Is a sample of 500 volunteers a reliable measure for a population of 2,500?

Solution:

No, even though the sample is large enough, the fact that the sample consists of volunteers makes it a self-selected sample, which is not reliable.

Exercise:

Problem:

A question on a survey reads: "Do you prefer the delicious taste of Brand X or the taste of Brand Y?" Is this a fair question?

Exercise:

Problem: Is a sample size of two representative of a population of five?

Solution:

No, even though the sample is a large portion of the population, two responses are not enough to justify any conclusions. Because the population is so small, it would be better to include everyone in the population to get the most accurate data.

Exercise:

Problem:

Is it possible for two experiments to be well run with similar sample sizes to get different data?

HOMEWORK

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or qualitative), and give an example of the data.

Exercise:

Problem: number of tickets sold to a concert

Solution:

quantitative discrete, 150

Exercise:

Problem: percent of body fat

Exercise:

Problem: favorite baseball team

Solution:

qualitative, Oakland A’s

Exercise:

Problem: time in line to buy groceries

Exercise:

Problem: number of students enrolled at Evergreen Valley College

---

Solution:

quantitative discrete, 11,234 students

Exercise:

Problem: most-watched television show

Exercise:

Problem: brand of toothpaste

Solution:

qualitative, Crest

Exercise:

Problem: distance to the closest movie theatre

Exercise:

Problem: age of executives in Fortune 500 companies

Solution:

quantitative continuous, 47.3 years

Exercise:

Problem: number of competing computer spreadsheet software packages

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every 8th house in the neighborhood around the park was interviewed.

Exercise:

Problem: “Number of times per week” is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Solution:

b

Exercise:

Problem: “Duration (amount of time)” is what type of data?

a. qualitative
b. quantitative discrete
c. quantitative continuous

Exercise:

Problem:

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Boston to Salt Lake City to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

a. Using complete sentences, list three things wrong with the way the survey was conducted.

b. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

Solution:

a. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.
b. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

Exercise:

Problem:

Suppose you want to determine the mean number of students per statistics class in your state. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Exercise:

Problem:

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Solution:

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

Exercise:

Problem:

List some practical difficulties involved in getting accurate results from a telephone survey.

Exercise:

Problem:

List some practical difficulties involved in getting accurate results from a mailed survey.

Solution:

Answers will vary. Sample Answer: Many people will not respond to mail surveys. If they do respond to the surveys, you can’t be sure who is responding. In addition, mailing lists can be incomplete.

Exercise:

Problem:

With your classmates, brainstorm some ways you could overcome these problems if you needed to conduct a phone or mail survey.

Exercise:

Problem:

The instructor takes her sample by gathering data on five randomly selected students from each Lake Tahoe Community College math class. The type of sampling she used is

a. cluster sampling
b. stratified sampling
c. simple random sampling
d. convenience sampling

Solution:

b

Exercise:

Problem:

A study was done to determine the age, number of times per week, and the duration (amount of time) of residents using a local park in San Jose. The first house in the neighborhood around the park was selected randomly and then every eighth house in the neighborhood around the park was interviewed. The sampling method was:

a. simple random
b. systematic
c. stratified
d. cluster

Exercise:

Problem:

Name the sampling method used in each of the following situations:

a. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.
b. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
c. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
d. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.

e. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

Solution:

convenience cluster stratified systematic simple random

Exercise:

Problem:

A "random survey" was conducted of 3,274 people of the "microprocessor generation" (people born since 1971, the year the microprocessor was invented). It was reported that 48% of those individuals surveyed stated that if they had $2,000 to spend, they would use it for computer equipment. Also, 66% of those surveyed considered themselves relatively savvy computer users.

a. Do you consider the sample size large enough for a study of this type? Why or why not?
b. Based on your "gut feeling," do you believe the percents accurately reflect the U.S. population for those individuals born since 1971? If not, do you think the percents of the population are actually higher or lower than the sample statistics? Why?
   Additional information: The survey, reported by Intel Corporation, was filled out by individuals who visited the Los Angeles Convention Center to see the Smithsonian Institute's road show called "America's Smithsonian."
c. With this additional information, do you feel that all demographic and ethnic groups were equally represented at the event? Why or why not?
d. With the additional information, comment on how accurately you think the sample statistics reflect the population parameters.

Exercise:

Problem:

The Well-Being Index is a survey that follows trends of U.S. residents on a regular basis. There are six areas of health and wellness covered in the survey: Life Evaluation, Emotional Health, Physical Health, Healthy Behavior, Work Environment, and Basic Access. Some of the questions used to measure the Index are listed below.

Identify the type of data obtained from each question used in this survey: qualitative, quantitative discrete, or quantitative continuous.

a. Do you have any health problems that prevent you from doing any of the things people your age can normally do?
b. During the past 30 days, for about how many days did poor health keep you from doing your usual activities?
c. In the last seven days, on how many days did you exercise for 30 minutes or more?
d. Do you have health insurance coverage?

Solution:

a. qualitative
b. quantitative discrete
c. quantitative discrete
d. qualitative

Exercise:

Problem:

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

a. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
b. What effect does the low response rate have on the reliability of the sample?
c. Are these problems examples of sampling error or nonsampling error?
d. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. These researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

Exercise:

Problem:

Crime-related and demographic statistics for 47 US states in 1960 were collected from government agencies, including the FBI's *Uniform Crime Report*. One analysis of this data found a strong connection between education and crime indicating that higher levels of education in a community correspond to higher crime rates.

Which of the potential problems with samples discussed in [link] could explain this connection?

Solution:

Causality: The fact that two variables are related does not guarantee that one variable is influencing the other. We cannot assume that crime rate impacts education level or that education level impacts crime rate.

Confounding: There are many factors that define a community other than education level and crime rate. Communities with high crime rates and high education levels may have other lurking variables that distinguish them from communities with lower crime rates and lower education levels. Because we cannot isolate these variables of interest, we cannot draw valid conclusions about the connection between education and

crime. Possible lurking variables include police expenditures, unemployment levels, region, average age, and size.

Exercise:

Problem:

YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

“Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?”[footnote]
lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).

As of April 25, 11 people responded to this question. Each participant answered “NO!”

Which of the potential problems with samples discussed in this module could explain this connection?

Exercise:

Problem:

A scholarly article about response rates begins with the following quote:

“Declining contact and cooperation rates in random digit dial (RDD) national telephone surveys raise serious concerns about the validity of estimates drawn from such research.”[footnote]
Scott Keeter et al., “Gauging the Impact of Growing Nonresponse on Estimates from a National RDD Telephone Survey,” Public Opinion Quarterly 70 no. 5 (2006), http://poq.oxfordjournals.org/content/70/5/759.full (accessed May 1, 2013).

The Pew Research Center for People and the Press admits:

“The percentage of people we interview – out of all we try to interview – has been declining over the past decade or more.”[footnote]

Frequently Asked Questions, Pew Research Center for the People & the Press, http://www.people-press.org/methodology/frequently-asked-questions/#dont-you-have-trouble-getting-people-to-answer-your-polls (accessed May 1, 2013).

a. What are some reasons for the decline in response rate over the past decade?
b. Explain why researchers are concerned with the impact of the declining response rate on public opinion polls.

Solution:

a. Possible reasons: increased use of caller id, decreased use of landlines, increased use of private numbers, voice mail, privacy managers, hectic nature of personal schedules, decreased willingness to be interviewed
b. When a large number of people refuse to participate, then the sample may not have the same characteristics of the population. Perhaps the majority of people willing to participate are doing so because they feel strongly about the subject of the survey.

Bringing It Together

Exercise:

Problem:

Seven hundred and seventy-one distance learning students at Long Beach City College responded to surveys in the 2010-11 academic year. Highlights of the summary report are listed in [link].

Have computer at home	96%

Unable to come to campus for classes	65%
Age 41 or over	24%
Would like LBCC to offer more DL courses	95%
Took DL classes due to a disability	17%
Live at least 16 miles from campus	13%
Took DL courses to fulfill transfer requirements	71%

LBCC Distance Learning Survey Results

a. What percent of the students surveyed do not have a computer at home?
b. About how many students in the survey live at least 16 miles from campus?
c. If the same survey were done at Great Basin College in Elko, Nevada, do you think the percentages would be the same? Why?

Exercise:

Problem:

Several online textbook retailers advertise that they have lower prices than on-campus bookstores. However, an important factor is whether the Internet retailers actually have the textbooks that students need in stock. Students need to be able to get textbooks promptly at the beginning of the college term. If the book is not available, then a student would not be able to get the textbook at all, or might get a delayed delivery if the book is back ordered.

A college newspaper reporter is investigating textbook availability at online retailers. He decides to investigate one textbook for each of the following seven subjects: calculus, biology, chemistry, physics, statistics, geology, and general engineering. He consults textbook industry sales

data and selects the most popular nationally used textbook in each of these subjects. He visits websites for a random sample of major online textbook sellers and looks up each of these seven textbooks to see if they are available in stock for quick delivery through these retailers. Based on his investigation, he writes an article in which he draws conclusions about the overall availability of all college textbooks through online textbook retailers.

Write an analysis of his study that addresses the following issues: Is his sample representative of the population of all college textbooks? Explain why or why not. Describe some possible sources of bias in this study, and how it might affect the results of the study. Give some suggestions about what could be done to improve the study.

Solution:

Answers will vary. Sample answer: The sample is not representative of the population of all college textbooks. Two reasons why it is not representative are that he only sampled seven subjects and he only investigated one textbook in each subject. There are several possible sources of bias in the study. The seven subjects that he investigated are all in mathematics and the sciences; there are many subjects in the humanities, social sciences, and other subject areas, (for example: literature, art, history, psychology, sociology, business) that he did not investigate at all. It may be that different subject areas exhibit different patterns of textbook availability, but his sample would not detect such results.

He also looked only at the most popular textbook in each of the subjects he investigated. The availability of the most popular textbooks may differ from the availability of other textbooks in one of two ways:

- the most popular textbooks may be more readily available online, because more new copies are printed, and more students nationwide are selling back their used copies OR
- the most popular textbooks may be harder to find available online, because more student demand exhausts the supply more quickly.

In reality, many college students do not use the most popular textbook in their subject, and this study gives no useful information about the situation for those less popular textbooks.

He could improve this study by:

- expanding the selection of subjects he investigates so that it is more representative of all subjects studied by college students, and
- expanding the selection of textbooks he investigates within each subject to include a mixed representation of both the most popular and less popular textbooks.

Glossary

Cluster Sampling
: a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable
: a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Convenience Sampling
: a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Discrete Random Variable
: a random variable (RV) whose outcomes are counted

Nonsampling Error
: an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Qualitative Data
: See Data.

Quantitative Data

See Data.

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias

not all members of the population are equally likely to be selected

Sampling Error

the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement

Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement

A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling

a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These randomly selected labels identify the members of your sample.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Frequency, Frequency Tables, and Levels of Measurement

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. However, when calculating the frequency, you may need to round your answers so that they are as precise as possible.

Answers and Rounding Off

A simple way to round off answers is to carry your final answer one more decimal place than was present in the original data. Round off only the final answer. Do not round off any intermediate results, if possible. If it becomes necessary to round off intermediate results, carry them to at least twice as many decimal places as the final answer. For example, the average of the three quiz scores four, six, and nine is 6.3, rounded off to the nearest tenth, because the data are whole numbers. Most answers will be rounded off in this manner.

It is not necessary to reduce most fractions in this course. Especially in Probability Topics, the chapter on probability, it is more helpful to leave an answer as an unreduced fraction.

Levels of Measurement

The way a set of data is measured is called its **level of measurement**. Correct statistical procedures depend on a researcher being familiar with levels of measurement. Not every statistical operation can be used with every set of data. Data can be classified into four levels of measurement. They are (from lowest to highest level):

- **Nominal scale level**
- **Ordinal scale level**
- **Interval scale level**
- **Ratio scale level**

Data that is measured using a **nominal scale** is **qualitative(categorical)**. Categories, colors, names, labels and favorite foods along with yes or no responses are examples of nominal level data. Nominal scale data are not ordered. For example, trying to classify people according to their favorite food does not make any sense. Putting pizza first and sushi second is not meaningful.

Smartphone companies are another example of nominal scale data. The data are the names of the companies that make smartphones, but there is no agreed upon order of these brands, even though people may have personal preferences. Nominal scale data cannot be used in calculations.

Data that is measured using an **ordinal scale** is similar to nominal scale data but there is a big difference. The ordinal scale data can be ordered. An example of ordinal scale data is a list of the top five national parks in the United States. The top five national parks in the United States can be ranked from one to five but we cannot measure differences between the data.

Another example of using the ordinal scale is a cruise survey where the responses to questions about the cruise are “excellent,” “good,” “satisfactory,” and “unsatisfactory.” These responses are ordered from the most desired response to the least desired. But the differences between two pieces

of data cannot be measured. Like the nominal scale data, ordinal scale data cannot be used in calculations.

Data that is measured using the **interval scale** is similar to ordinal level data because it has a definite ordering but there is a difference between data. The differences between interval scale data can be measured though the data does not have a starting point.

Temperature scales like Celsius (C) and Fahrenheit (F) are measured by using the interval scale. In both temperature measurements, 40° is equal to 100° minus 60°. Differences make sense. But 0 degrees does not because, in both scales, 0 is not the absolute lowest temperature. Temperatures like -10° F and -15° C exist and are colder than 0.

Interval level data can be used in calculations, but one type of comparison cannot be done. 80° C is not four times as hot as 20° C (nor is 80° F four times as hot as 20° F). There is no meaning to the ratio of 80 to 20 (or four to one).

Data that is measured using the **ratio scale** takes care of the ratio problem and gives you the most information. Ratio scale data is like interval scale data, but it has a 0 point and ratios can be calculated. For example, four multiple choice statistics final exam scores are 80, 68, 20 and 92 (out of a possible 100 points). The exams are machine-graded.

The data can be put in order from lowest to highest: 20, 68, 80, 92.

The differences between the data have meaning. The score 92 is more than the score 68 by 24 points. Ratios can be calculated. The smallest score is 0. So 80 is four times 20. The score of 80 is four times better than the score of 20.

Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 56332475235654435253.

[link] lists the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6

DATA VALUE	FREQUENCY
6	2
7	1

Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to [link], there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15
3	5	$\frac{5}{20}$ or 0.25
4	3	$\frac{3}{20}$ or 0.15
5	6	$\frac{6}{20}$ or 0.30
6	2	$\frac{2}{20}$ or 0.10
7	1	$\frac{1}{20}$ or 0.05

Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of [link] is $\frac{20}{20}$, or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in [link].

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

Note:
NOTE
Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

[link] represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
59.95–61.95	5	$\frac{5}{100}$ = 0.05	0.05
61.95–63.95	3	$\frac{3}{100}$ = 0.03	0.05 + 0.03 = 0.08

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
63.95–65.95	15	$\frac{15}{100}$ = 0.15	0.08 + 0.15 = 0.23
65.95–67.95	40	$\frac{40}{100}$ = 0.40	0.23 + 0.40 = 0.63
67.95–69.95	17	$\frac{17}{100}$ = 0.17	0.63 + 0.17 = 0.80
69.95–71.95	12	$\frac{12}{100}$ = 0.12	0.80 + 0.12 = 0.92
71.95–73.95	7	$\frac{7}{100}$ = 0.07	0.92 + 0.07 = 0.99
73.95–75.95	1	$\frac{1}{100}$ = 0.01	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 59.95 to 61.95 inches
- 61.95 to 63.95 inches
- 63.95 to 65.95 inches
- 65.95 to 67.95 inches
- 67.95 to 69.95 inches
- 69.95 to 71.95 inches
- 71.95 to 73.95 inches
- 73.95 to 75.95 inches

Note:
Note
This example is used again in Descriptive Statistics, where the method used to compute the intervals will be explained.

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

Example:
Exercise:

Problem: From [link], find the percentage of heights that are less than 65.95 inches.

Solution:

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are 5 + 3 + 15 = 23 players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then $\frac{23}{100}$ or 23%. This percentage is the cumulative relative frequency entry in the third row.

Note:
Try It
Exercise:

Problem: [link] shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
2.95–4.97	6	$\frac{6}{50}$ = 0.12	0.12
4.97–6.99	7	$\frac{7}{50}$ = 0.14	0.12 + 0.14 = 0.26
6.99–9.01	15	$\frac{15}{50}$ = 0.30	0.26 + 0.30 = 0.56
9.01–11.03	8	$\frac{8}{50}$ = 0.16	0.56 + 0.16 = 0.72
11.03–13.05	9	$\frac{9}{50}$ = 0.18	0.72 + 0.18 = 0.90
13.05–15.07	5	$\frac{5}{50}$ = 0.10	0.90 + 0.10 = 1.00
	Total = 50	Total = 1.00	

From [link], find the percentage of rainfall that is less than 9.01 inches.

Solution:
Try It Solutions

0.56 or 56%

Example:
Exercise:

Problem:

From [link], find the percentage of heights that fall between 61.95 and 65.95 inches.

Solution:

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Note:
Try It
Exercise:

Problem: From [link], find the percentage of rainfall that is between 6.99 and 13.05 inches.

Solution:
Try It Solutions

0.30 + 0.16 + 0.18 = 0.64 or 64%

Example:
Exercise:

Problem:

Use the heights of the 100 male semiprofessional soccer players in [link]. Fill in the blanks and check your answers.

a. The percentage of heights that are from 67.95 to 71.95 inches is: ____.
b. The percentage of heights that are from 67.95 to 73.95 inches is: ____.
c. The percentage of heights that are more than 65.95 inches is: ____.
d. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ____.
e. What kind of data are the heights?
f. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the

previous relative frequencies to the relative frequency for the current row.

Solution:

a. 29%
b. 36%
c. 77%
d. 87
e. quantitative continuous
f. get rosters from each team and choose a simple random sample from each

Note:
Try It
Exercise:

Problem:

From [link], find the number of towns that have rainfall between 2.95 and 9.01 inches.

Solution:
Try It Solutions

6 + 7 + 15 = 28 towns

Note:
Collaborative Exercise
In your class, have someone conduct a survey of the number of siblings (brothers and sisters) each student has. Create a frequency table. Add to it a relative frequency column and a cumulative relative frequency column. Answer the following questions:

1. What percentage of the students in your class have no siblings?
2. What percentage of the students have from one to three siblings?
3. What percentage of the students have fewer than three siblings?

Example:
Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2 5 7 3 2 10 18 15 20 7 10 18 5 12 13 12 4 5 10. [link] was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	$\frac{3}{19}$	0.1579
4	1	$\frac{1}{19}$	0.2105
5	3	$\frac{3}{19}$	0.1579
7	2	$\frac{2}{19}$	0.2632
10	3	$\frac{4}{19}$	0.4737
12	2	$\frac{2}{19}$	0.7895
13	1	$\frac{1}{19}$	0.8421
15	1	$\frac{1}{19}$	0.8948
18	1	$\frac{1}{19}$	0.9474
20	1	$\frac{1}{19}$	1.0000

Frequency of Commuting Distances

Exercise:

Problem:

a. Is the table correct? If it is not correct, what is wrong?
b. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
c. What fraction of the people surveyed commute five or seven miles?
d. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

Solution:

a. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
b. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.
c. $\frac{5}{19}$

d. $\frac{7}{19}$, $\frac{12}{19}$, $\frac{7}{19}$

Note:
Try It
Exercise:

Problem:

[link] represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get between 11.03 and 13.05 inches of rainfall each year?

Solution:
Try It Solutions

$\frac{9}{50}$

Example:
[link] contains the total number of deaths worldwide as a result of earthquakes for the period from 2000 to 2012.

Year	Total Number of Deaths
2000	231
2001	21,357
2002	11,685
2003	33,819
2004	228,802
2005	88,003
2006	6,605
2007	712

Year	Total Number of Deaths
2008	88,011
2009	1,790
2010	320,120
2011	21,953
2012	768
Total	823,856

Exercise:

Problem: Answer the following questions.

a. What is the frequency of deaths measured from 2006 through 2009?
b. What percentage of deaths occurred after 2009?
c. What is the relative frequency of deaths that occurred in 2003 or earlier?
d. What is the percentage of deaths that occurred in 2004?
e. What kind of data are the numbers of deaths?
f. The Richter scale is used to quantify the energy produced by an earthquake. Examples of Richter scale numbers are 2.3, 4.0, 6.1, and 7.0. What kind of data are these numbers?

Solution:

a. 97,118 (11.8%)
b. 41.6%
c. 67,092/823,356 or 0.081 or 8.1 %
d. 27.8%
e. Quantitative discrete
f. Quantitative continuous

Note:
Try It
Exercise:

Problem:

[link] contains the total number of fatal motor vehicle traffic crashes in the United States for the period from 1994 to 2011.

Year	Total Number of Crashes	Year	Total Number of Crashes
1994	36,254	2004	38,444
1995	37,241	2005	39,252
1996	37,494	2006	38,648
1997	37,324	2007	37,435
1998	37,107	2008	34,172
1999	37,140	2009	30,862
2000	37,526	2010	30,296
2001	37,862	2011	29,757
2002	38,491	Total	653,782
2003	38,477		

Answer the following questions.

a. What is the frequency of deaths measured from 2000 through 2004?
b. What percentage of deaths occurred after 2006?
c. What is the relative frequency of deaths that occurred in 2000 or before?
d. What is the percentage of deaths that occurred in 2011?
e. What is the cumulative relative frequency for 2006? Explain what this number tells you about the data.

Solution:
Try It Solutions

a. 190,800 (29.2%)
b. 24.9%
c. 260,086/653,782 or 39.8%
d. 4.6%
e. 75.1% of all fatal traffic crashes for the period from 1994 to 2011 happened from 1994 to 2006.

References

"State & County QuickFacts," U.S. Census Bureau. http://quickfacts.census.gov/qfd/download_data.html (accessed May 1, 2013).

"State & County QuickFacts: Quick, easy access to facts about people, business, and geography," U.S. Census Bureau. http://quickfacts.census.gov/qfd/index.html (accessed May 1, 2013).

"Table 5: Direct hits by mainland United States Hurricanes (1851-2004)," National Hurricane Center, http://www.nhc.noaa.gov/gifs/table5.gif (accessed May 1, 2013).

"Levels of Measurement," http://infinity.cos.edu/faculty/woodbury/stats/tutorial/Data_Levels.htm (accessed May 1, 2013).

Courtney Taylor, "Levels of Measurement," about.com, http://statistics.about.com/od/HelpandTutorials/a/Levels-Of-Measurement.htm (accessed May 1, 2013).

David Lane. "Levels of Measurement," Connexions, http://cnx.org/content/m10809/latest/ (accessed May 1, 2013).

Chapter Review

Some calculations generate numbers that are artificially precise. It is not necessary to report a value to eight decimal places when the measures that generated that value were only accurate to the nearest tenth. Round off your final answer to one more decimal place than was present in the original data. This means that if you have data measured to the nearest tenth of a unit, report the final statistic to the nearest hundredth.

In addition to rounding your answers, you can measure your data using the following four levels of measurement.

- **Nominal scale level:** data that cannot be ordered nor can it be used in calculations
- **Ordinal scale level:** data that can be ordered; the differences cannot be measured
- **Interval scale level:** data with a definite ordering but no starting point; the differences can be measured, but there is no such thing as a ratio.
- **Ratio scale level:** data with a starting point that can be ordered; the differences have meaning and ratios can be calculated.

When organizing data, it is important to know how many times a value appears. How many statistics students study five hours or more for an exam? What percent of families on our block own two pets? Frequency, relative frequency, and cumulative relative frequency are measures that answer questions like these.
Exercise:

Problem: What type of measure scale is being used? Nominal, ordinal, interval or ratio.

a. High school soccer players classified by their athletic ability: Superior, Average, Above average
b. Baking temperatures for various main dishes: 350, 400, 325, 250, 300
c. The colors of crayons in a 24-crayon box

d. Social security numbers
e. Incomes measured in dollars
f. A satisfaction survey of a social website by number: 1 = very satisfied, 2 = somewhat satisfied, 3 = not satisfied
g. Political outlook: extreme left, left-of-center, right-of-center, extreme right
h. Time of day on an analog watch
i. The distance in miles to the closest grocery store
j. The dates 1066, 1492, 1644, 1947, and 1944
k. The heights of 21–65 year-old women
l. Common letter grades: A, B, C, D, and F

Solution:

a. ordinal
b. interval
c. nominal
d. nominal
e. ratio
f. ordinal
g. nominal
h. interval
i. ratio
j. interval
k. ratio
l. ordinal

HOMEWORK

Exercise:

Problem:

Fifty part-time students were asked how many courses they were taking this term. The (incomplete) results are shown below:

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
1	30	0.6	
2	15		

# of Courses	Frequency	Relative Frequency	Cumulative Relative Frequency
3			

Part-time Student Course Loads

a. Fill in the blanks in [link].
b. What percent of students take exactly two courses?
c. What percent of students take one or two courses?

Exercise:

Problem:

Sixty adults with gum disease were asked the number of times per week they used to floss before their diagnosis. The (incomplete) results are shown in [link].

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Freq.
0	27	0.4500	
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

Flossing Frequency for Adults with Gum Disease

a. Fill in the blanks in [link].
b. What percent of adults flossed six times per week?
c. What percent flossed at most three times per week?

Solution:

a.

# Flossing per Week	Frequency	Relative Frequency	Cumulative Relative Frequency
0	27	0.4500	0.4500
1	18	0.3000	0.7500
3	11	0.1833	0.9333
6	3	0.0500	0.9833
7	1	0.0167	1

b. 5.00%
c. 93.33%

Exercise:

Problem:

Nineteen immigrants to the U.S were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows: 2 5 7 2 2 10 20 15 0 7 0 20 5 12 15 12 4 5 10 .

[link] was produced.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
0	2	$\frac{2}{19}$	0.1053
2	3	$\frac{3}{19}$	0.2632
4	1	$\frac{1}{19}$	0.3158
5	3	$\frac{3}{19}$	0.4737
7	2	$\frac{2}{19}$	0.5789
10	2	$\frac{2}{19}$	0.6842
12	2	$\frac{2}{19}$	0.7895

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
15	1	$\frac{1}{19}$	0.8421
20	1	$\frac{1}{19}$	1.0000

Frequency of Immigrant Survey Responses

a. Fix the errors in [link]. Also, explain how someone might have arrived at the incorrect number(s).
b. Explain what is wrong with this statement: "47 percent of the people surveyed have lived in the U.S. for 5 years."
c. Fix the statement in **b** to make it correct.
d. What fraction of the people surveyed have lived in the U.S. five or seven years?
e. What fraction of the people surveyed have lived in the U.S. at most 12 years?
f. What fraction of the people surveyed have lived in the U.S. fewer than 12 years?
g. What fraction of the people surveyed have lived in the U.S. from five to 20 years, inclusive?

Exercise:

Problem:

How much time does it take to travel to work? [link] shows the mean commute time by state for workers at least 16 years old who are not working at home. Find the mean travel time, and round off the answer properly.

24.0	24.3	25.9	18.9	27.5	17.9	21.8	20.9	16.7	27.3
18.2	24.7	20.0	22.6	23.9	18.0	31.4	22.3	24.0	25.5
24.7	24.6	28.1	24.9	22.6	23.6	23.4	25.7	24.8	25.5
21.2	25.7	23.1	23.0	23.9	26.0	16.3	23.1	21.4	21.5
27.0	27.0	18.6	31.7	23.3	30.1	22.9	23.3	21.7	18.6

Solution:

The sum of the travel times is 1,173.1. Divide the sum by 50 to calculate the mean value: 23.462. Because each state's travel time was measured to the nearest tenth, round this calculation to the nearest hundredth: 23.46.

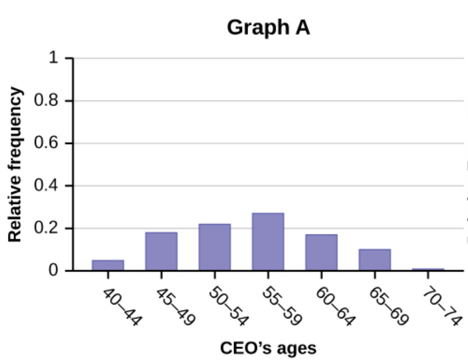
Exercise:

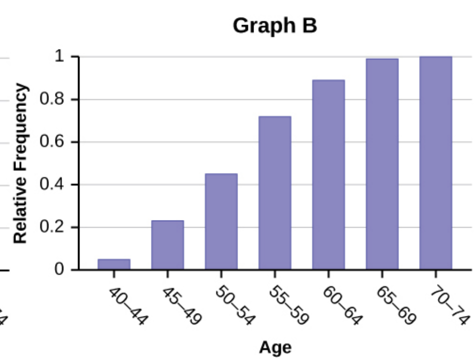
Problem:

Forbes magazine published data on the best small firms in 2012. These were firms which had been publicly traded for at least a year, have a stock price of at least $5 per share, and have reported annual revenue between $5 million and $1 billion. [link] shows the ages of the chief executive officers for the first 60 ranked firms.

Age	Frequency	Relative Frequency	Cumulative Relative Frequency
40–44	3		
45–49	11		
50–54	13		
55–59	16		
60–64	10		
65–69	6		
70–74	1		

a. What is the frequency for CEO ages between 54 and 65?
b. What percentage of CEOs are 65 years or older?
c. What is the relative frequency of ages under 50?
d. What is the cumulative relative frequency for CEOs younger than 55?
e. Which graph shows the relative frequency and which shows the cumulative relative frequency?





Use the following information to answer the next two exercises: [link] contains data on hurricanes that have made direct hits on the U.S. Between 1851 and 2004. A hurricane is given a strength category rating based on the minimum wind speed generated by the storm.

Category	**Number of Direct Hits**	**Relative Frequency**	**Cumulative Frequency**
1	109	0.3993	0.3993
2	72	0.2637	0.6630
3	71	0.2601	
4	18		0.9890
5	3	0.0110	1.0000
	Total = 273		

Frequency of Hurricane Direct Hits

Exercise:

Problem: What is the relative frequency of direct hits that were category 4 hurricanes?

a. 0.0768
b. 0.0659
c. 0.2601
d. Not enough information to calculate

Solution:

b

Exercise:

Problem:

What is the relative frequency of direct hits that were AT MOST a category 3 storm?

a. 0.3480
b. 0.9231
c. 0.2601
d. 0.3370

Glossary

Cumulative Relative Frequency

The term applies to an ordered set of observations from smallest to largest. The cumulative relative frequency is the sum of the relative frequencies for all values that are less than or equal to the given value.

Frequency

the number of times a value of the data occurs

Relative Frequency

the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes to the total number of outcomes

Derived copy of Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **explanatory variable**. The affected variable is called the **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the

researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

*Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.*[footnote]
McClung, M. Collins, D. "Because I know it will!": placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

Example:
Exercise:

Problem:

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years, but he does not know whether he is taking aspirin or the placebo. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify the following values for this study: population, sample, experimental units, explanatory variable, response variable, treatments.

Solution:

The *population* is men aged 50 to 84.
The *sample* is the 400 men who participated.
The *experimental units* are the individual men in the study.
The *explanatory variable* is oral medication.
The *treatments* are aspirin and a placebo.
The *response variable* is whether a subject had a heart attack.

Example:
Exercise:

Problem:

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

a. Describe the explanatory and response variables in this study.
b. What are the treatments?
c. Identify any lurking variables that could interfere with this study.
d. Is it possible to use blinding in this study?

Solution:

a. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
b. There are two treatments: a floral-scented mask and an unscented mask.
c. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
d. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

Example:

Exercise:

Problem:

A researcher wants to study the effects of birth order on personality. Explain why this study could not be conducted as a randomized experiment. What is the main problem in a study that cannot be designed as a randomized experiment?

Solution:

The explanatory variable is birth order. You cannot randomly assign a person's birth order. Random assignment eliminates the impact of lurking variables. When you cannot assign subjects to treatment groups at random, there will be differences between the groups other than the explanatory variable.

Note:
Try It
Exercise:

Problem:

You are concerned about the effects of texting on driving performance. Design a study to test the response time of drivers while texting and while driving only. How many seconds does it take for a driver to respond when a leading car hits the brakes?

a. Describe the explanatory and response variables in the study.
b. What are the treatments?
c. What should you consider when selecting participants?
d. Your research partner wants to divide participants randomly into two groups: one to drive without distraction and one to text and drive simultaneously. Is this a good idea? Why or why not?
e. Identify any lurking variables that could interfere with this study.
f. How can blinding be used in this study?

Solution:
Try It Solutions

a. Explanatory: presence of distraction from texting; response: response time measured in seconds
b. Driving without distraction and driving while texting
c. Answers will vary. Possible responses: Do participants regularly send and receive text messages? How long has the subject been driving? What is the age of the participants? Do participants have similar texting and driving experience?
d. This is not a good plan because it compares drivers with different abilities. It would be better to assign both treatments to each participant in random order.
e. Possible responses include: texting ability, driving experience, type of phone.
f. The researchers observing the trials and recording response time could be blinded to the treatment being applied.

Ethics

The widespread misuse and misrepresentation of statistical information often gives the field a bad name. Some say that "numbers don't lie," but the people who use numbers to support their claims often do.

A recent investigation of famous social psychologist, Diederik Stapel, has led to the retraction of his articles from some of the world's top journals including *Journal of Experimental Social Psychology, Social Psychology, Basic and Applied Social Psychology, British Journal of Social Psychology,* and the magazine *Science*. Diederik Stapel is a former professor at Tilburg University in the Netherlands. Over the past two years, an extensive investigation involving three universities where Stapel has worked concluded that the psychologist is guilty of fraud on a colossal scale.

Falsified data taints over 55 papers he authored and 10 Ph.D. dissertations that he supervised.

Stapel did not deny that his deceit was driven by ambition. But it was more complicated than that, he told me. He insisted that he loved social psychology but had been frustrated by the messiness of experimental data, which rarely led to clear conclusions. His lifelong obsession with elegance and order, he said, led him to concoct sexy results that journals found attractive. "It was a quest for aesthetics, for beauty—instead of the truth," he said. He described his behavior as an addiction that drove him to carry out acts of increasingly daring fraud, like a junkie seeking a bigger and better high.[footnote]
Yudhijit Bhattacharjee, "The Mind of a Con Man," Magazine, New York Times, April 26, 2013. Available online at: http://www.nytimes.com/2013/04/28/magazine/diederik-stapels-audacious-academic-fraud.html?src=dayp&_r=2& (accessed May 1, 2013).

The committee investigating Stapel concluded that he is guilty of several practices including:

- creating datasets, which largely confirmed the prior expectations,
- altering data in existing datasets,
- changing measuring instruments without reporting the change, and
- misrepresenting the number of experimental subjects.

Clearly, it is never acceptable to falsify data the way this researcher did. Sometimes, however, violations of ethics are not as easy to spot.

Researchers have a responsibility to verify that proper methods are being followed. The report describing the investigation of Stapel's fraud states that, "statistical flaws frequently revealed a lack of familiarity with elementary statistics."[footnote] Many of Stapel's co-authors should have spotted irregularities in his data. Unfortunately, they did not know very much about statistical analysis, and they simply trusted that he was collecting and reporting data properly.
"Flawed Science: The Fraudulent Research Practices of Social Psychologist Diederik Stapel," Tillburg University, November 28, 2012, http://www.tilburguniversity.edu/upload/064a10cd-bce5-4385-b9ff-

05b840caeae6_120695_Rapp_nov_2012_UK_web.pdf (accessed May 1, 2013).

Many types of statistical fraud are difficult to spot. Some researchers simply stop collecting data once they have just enough to prove what they had hoped to prove. They don't want to take the chance that a more extensive study would complicate their lives by producing data contradicting their hypothesis.

Professional organizations, like the American Statistical Association, clearly define expectations for researchers. There are even laws in the federal code about the use of research data.

When a statistical study uses human participants, as in medical studies, both ethics and the law dictate that researchers should be mindful of the safety of their research subjects. The U.S. Department of Health and Human Services oversees federal regulations of research studies with the aim of protecting participants. When a university or other research institution engages in research, it must ensure the safety of all human subjects. For this reason, research institutions establish oversight committees known as **Institutional Review Boards (IRB)**. All planned studies must be approved in advance by the IRB. Key protections that are mandated by law include the following:

- Risks to participants must be minimized and reasonable with respect to projected benefits.
- Participants must give **informed consent**. This means that the risks of participation must be clearly explained to the subjects of the study. Subjects must consent in writing, and researchers are required to keep documentation of their consent.
- Data collected from individuals must be guarded carefully to protect their privacy.

These ideas may seem fundamental, but they can be very difficult to verify in practice. Is removing a participant's name from the data record sufficient to protect privacy? Perhaps the person's identity could be discovered from the data that remains. What happens if the study does not proceed as planned and risks arise that were not anticipated? When is informed consent really necessary? Suppose your doctor wants a blood sample to check your

cholesterol level. Once the sample has been tested, you expect the lab to dispose of the remaining blood. At that point the blood becomes biological waste. Does a researcher have the right to take it for use in a study?

It is important that students of statistics take time to consider the ethical questions that arise in statistical studies. How prevalent is fraud in statistical studies? You might be surprised—and disappointed. There is a website (www.retractionwatch.com) dedicated to cataloging retractions of study articles that have been proven fraudulent. A quick glance will show that the misuse of statistics is a bigger problem than most people realize.

Vigilance against fraud requires knowledge. Learning the basic theory of statistics will empower you to analyze statistical studies critically.

Example:
Exercise:

Problem:

Describe the unethical behavior in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A researcher is collecting data in a community.

a. She selects a block where she is comfortable walking because she knows many of the people living on the street.
b. No one seems to be home at four houses on her route. She does not record the addresses and does not return at a later time to try to find residents at home.
c. She skips four houses on her route because she is running late for an appointment. When she gets home, she fills in the forms by selecting random answers from other residents in the neighborhood.

Solution:

a. By selecting a convenient sample, the researcher is intentionally selecting a sample that could be biased. Claiming that this sample represents the community is misleading. The researcher needs to select areas in the community at random.
b. Intentionally omitting relevant data will create bias in the sample. Suppose the researcher is gathering information about jobs and child care. By ignoring people who are not home, she may be missing data from working families that are relevant to her study. She needs to make every effort to interview all members of the target sample.
c. It is never acceptable to fake data. Even though the responses she uses are “real” responses provided by other participants, the duplication is fraudulent and can create bias in the data. She needs to work diligently to interview everyone on her route.

Note:
Try It
Exercise:

Problem:

Describe the unethical behavior, if any, in each example and describe how it could impact the reliability of the resulting data. Explain how the problem should be corrected.

A study is commissioned to determine the favorite brand of fruit juice among teens in California.

a. The survey is commissioned by the seller of a popular brand of apple juice.
b. There are only two types of juice included in the study: apple juice and cranberry juice.
c. Researchers allow participants to see the brand of juice as samples are poured for a taste test.

d. Twenty-five percent of participants prefer Brand X, 33% prefer Brand Y and 42% have no preference between the two brands. Brand X references the study in a commercial saying “Most teens like Brand X as much as or more than Brand Y.”

Solution:

a. This is not necessarily a problem. The study should be monitored carefully, however, to ensure that the company is not pressuring researchers to return biased results.
b. If the researchers truly want to determine the favorite brand of juice, then researchers should ask teens to compare different brands of the same type of juice. Choosing a sweet juice to compare against a sharp-flavored juice will not lead to an accurate comparison of brand quality.
c. Participants could be biased by the knowledge. The results may be different from those obtained in a blind taste test.
d. The commercial tells the truth, but not the whole truth. It leads consumers to believe that Brand X was preferred by more participants than Brand Y while the opposite is true.

References

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/vitamin-e/ (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online

at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html (accessed May 1, 2013).

M.L. Jacskon et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

“Earthquake Information by Year,” U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year/ (accessed May 1, 2013).

“Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

“America’s Best Small Companies,” http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

“April 2013 Air Travel Consumer Report,” U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, “Statistics can be Misleading,” econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, “Ethics in Statistics,” Based on “Building an Ethics Module for Business, Science, and Engineering Students” by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).

Chapter Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”[footnote] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics1.pdf (accessed May 1, 2013).

Exercise:

Problem:

Design an experiment. Identify the explanatory and response variables. Describe the population being studied and the experimental units. Explain the treatments that will be used and how they will be assigned to the experimental units. Describe how blinding and placebos may be used to counter the power of suggestion.

Exercise:

Problem:

Discuss potential violations of the rule requiring informed consent.

a. Inmates in a correctional facility are offered good behavior credit in return for participation in a study.
b. A research study is designed to investigate a new children's allergy medication.
c. Participants in a study are told that the new medication being tested is highly promising, but they are not told that only a small portion of participants will receive the new medication. Others will receive placebo treatments and traditional treatments.

Solution:

a. Inmates may not feel comfortable refusing participation, or may feel obligated to take advantage of the promised benefits. They may not feel truly free to refuse participation.
b. Parents can provide consent on behalf of their children, but children are not competent to provide consent for themselves.
c. All risks and benefits must be clearly outlined. Study participants must be informed of relevant aspects of the study in order to give appropriate consent.

HOMEWORK

Exercise:

Problem:

How does sleep deprivation affect your ability to drive? A recent study measured the effects on 19 professional drivers. Each driver participated in two experimental sessions: one after normal sleep and one after 27 hours of total sleep deprivation. The treatments were assigned in random order. In each session, performance was measured on a variety of tasks including a driving simulation.

Use key terms from this module to describe the design of this experiment.

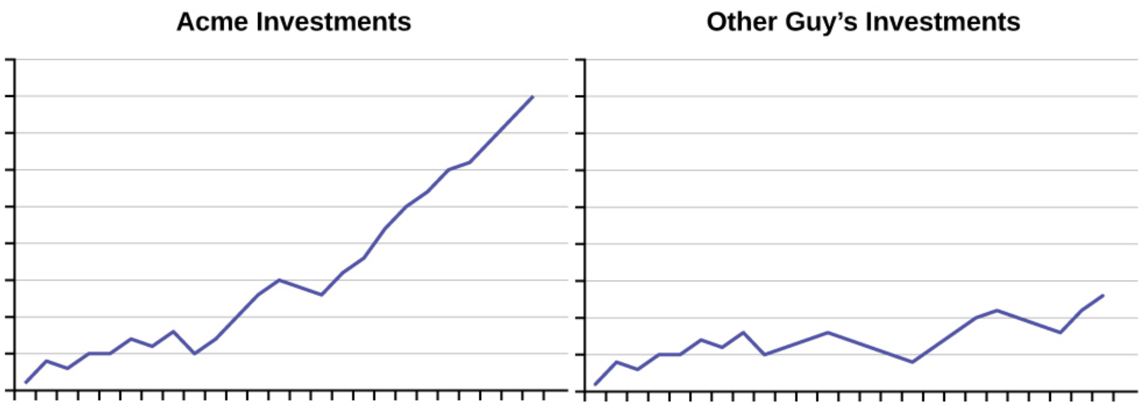
Solution:

Explanatory variable: amount of sleep
Response variable: performance measured in assigned tasks
Treatments: normal sleep and 27 hours of total sleep deprivation
Experimental Units: 19 professional drivers
Lurking variables: none – all drivers participated in both treatments
Random assignment: treatments were assigned in random order; this eliminated the effect of any “learning” that may take place during the first experimental session
Control/Placebo: completing the experimental session under normal sleep conditions
Blinding: researchers evaluating subjects’ performance must not know which treatment is being applied at the time

Exercise:

Problem:

An advertisement for Acme Investments displays the two graphs in [link] to show the value of Acme’s product in comparison with the Other Guy’s product. Describe the potentially misleading visual effect of these comparison graphs. How can this be corrected?

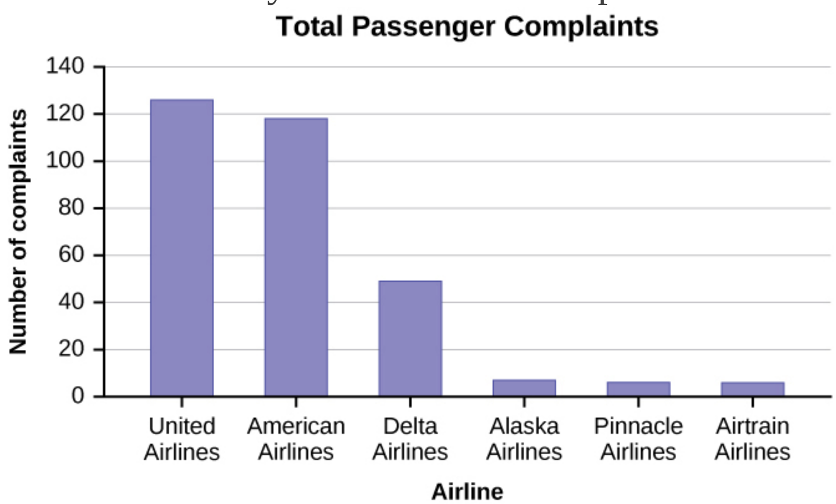


As the graphs show, Acme consistently outperforms the Other Guys!

Exercise:

Problem:

The graph in [link] shows the number of complaints for six different airlines as reported to the US Department of Transportation in February 2013. Alaska, Pinnacle, and Airtran Airlines have far fewer complaints reported than American, Delta, and United. Can we conclude that American, Delta, and United are the worst airline carriers since they have the most complaints?



Solution:

You cannot assume that the numbers of complaints reflect the quality of the airlines. The airlines shown with the greatest number of complaints are the ones with the most passengers. You must consider the appropriateness of methods for presenting data; in this case displaying totals is misleading.

Glossary

Explanatory Variable
: the independent variable in an experiment; the value controlled by researchers

Treatments
: different values or components of the explanatory variable applied in an experiment

Response Variable
: the dependent variable in an experiment; the value that is measured for change at the end of an experiment

Experimental Unit
: any individual or object to be measured

Lurking Variable
: a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Random Assignment
: the act of organizing experimental units into treatment groups using random methods

Control Group
: a group in a randomized experiment that receives an inactive treatment but is otherwise managed exactly as the other groups

Informed Consent

Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board
: a committee tasked with oversight of research programs that involve human subjects

Placebo
: an inactive treatment that has no real effect on the explanatory variable

Blinding
: not telling participants which treatment a subject is receiving

Double-blinding
: the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Stats Lab 1

Note:
Data Collection Experiment
Class Time:
Names:
Student Learning Outcomes

- The student will demonstrate the systematic sampling technique.
- The student will construct relative frequency tables.
- The student will interpret results and their differences from different data groupings.

Movie Survey
Ask five classmates from a different class how many movies they saw at the theater last month. Do not include rented movies.

1. Record the data.
2. In class, randomly pick one person. On the class list, mark that person's name. Move down four names on the class list. Mark that person's name. Continue doing this until you have marked 12 names. You may need to go back to the start of the list. For each marked name record the five data values. You now have a total of 60 data values.
3. For each name marked, record the data.

______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______
______	______	______	______	______	______	______	______

Order the Data
Complete the two relative frequency tables below using your class data.

Number of Movies	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0			
1			
2			
3			
4			
5			
6			
7+			

Frequency of Number of Movies Viewed

Number of Movies	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0–1			
2–3			
4–5			
6–7+			

Frequency of Number of Movies Viewed

1. Using the tables, find the percent of data that is at most two. Which table did you use and why?
2. Using the tables, find the percent of data that is at most three. Which table did you use and why?
3. Using the tables, find the percent of data that is more than two. Which table did you use and why?
4. Using the tables, find the percent of data that is more than three. Which table did you use and why?

Discussion Questions

1. Is one of the tables “more correct” than the other? Why or why not?
2. In general, how could you group the data differently? Are there any advantages to either way of grouping the data?
3. Why did you switch between tables, if you did, when answering the question above?

Stats Lab 2

Note:
Sampling Experiment
Class Time:
Names:
Student Learning Outcomes

- The student will demonstrate the simple random, systematic, stratified, and cluster sampling techniques.
- The student will explain the details of each procedure used.

In this lab, you will be asked to pick several random samples of restaurants. In each case, describe your procedure briefly, including how you might have used the random number generator, and then list the restaurants in the sample you obtained.

Note:
Note
The following section contains restaurants stratified by city into columns and grouped horizontally by entree cost (clusters).

Restaurants Stratified by City and Entree Cost

Entree Cost	**Under $10**	**$10 to under $15**	**$15 to under $20**	**Over $20**

Entree Cost	Under $10	$10 to under $15	$15 to under $20	Over $20
San Jose	El Abuelo Taq, Pasta Mia, Emma's Express, Bamboo Hut	Emperor's Guard, Creekside Inn	Agenda, Gervais, Miro's	Blake's, Eulipia, Hayes Mansion, Germania
Palo Alto	Senor Taco, Olive Garden, Taxi's	Ming's, P.A. Joe's, Stickney's	Scott's Seafood, Poolside Grill, Fish Market	Sundance Mine, Maddalena's, Spago's
Los Gatos	Mary's Patio, Mount Everest, Sweet Pea's, Andele Taqueria	Lindsey's, Willow Street	Toll House	Charter House, La Maison Du Cafe
Mountain View	Maharaja, New Ma's, Thai-Rific, Garden Fresh	Amber Indian, La Fiesta, Fiesta del Mar, Dawit	Austin's, Shiva's, Mazeh	Le Petit Bistro
Cupertino	Hobees, Hung Fu, Samrat, Panda Express	Santa Barb. Grill, Mand. Gourmet, Bombay Oven, Kathmandu West	Fontana's, Blue Pheasant	Hamasushi, Helios

Entree Cost	Under $10	$10 to under $15	$15 to under $20	Over $20
Sunnyvale	Chekijababi, Taj India, Full Throttle, Tia Juana, Lemon Grass	Pacific Fresh, Charley Brown's, Cafe Cameroon, Faz, Aruba's	Lion & Compass, The Palace, Beau Sejour	
Santa Clara	Rangoli, Armadillo Willy's, Thai Pepper, Pasand	Arthur's, Katie's Cafe, Pedro's, La Galleria	Birk's, Truya Sushi, Valley Plaza	Lakeside, Mariani's

Restaurants Used in Sample

A Simple Random Sample
Pick a **simple random sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________	11. __________
2. __________	7. __________	12. __________
3. __________	8. __________	13. __________
4. __________	9. __________	14. __________
5. __________	10. __________	15. __________

A Systematic Sample
Pick a **systematic sample** of 15 restaurants.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________	11. __________
2. __________	7. __________	12. __________
3. __________	8. __________	13. __________
4. __________	9. __________	14. __________
5. __________	10. __________	15. __________

A Stratified Sample
Pick a **stratified sample**, by city, of 20 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________	11. __________	16. __________
2. __________	7. __________	12. __________	17. __________
3. __________	8. __________	13. __________	18. __________

4. __________	9. __________	14. __________	19. __________
5. __________	10. __________	15. __________	20. __________

A Stratified Sample

Pick a **stratified sample**, by entree cost, of 21 restaurants. Use 25% of the restaurants from each stratum. Round to the nearest whole number.

1. Describe your procedure.
2. Complete the table with your sample.

1. __________	6. __________	11. __________	16. __________
2. __________	7. __________	12. __________	17. __________
3. __________	8. __________	13. __________	18. __________
4. __________	9. __________	14. __________	19. __________
5. __________	10. __________	15. __________	20. __________
			21. __________

A Cluster Sample

Pick a **cluster sample** of restaurants from two cities. The number of restaurants will vary.

1. Describe your procedure.

2. Complete the table with your sample.

1. __________	6. __________	11. __________	16. __________	21. __________
2. __________	7. __________	12. __________	17. __________	22. __________
3. __________	8. __________	13. __________	18. __________	23. __________
4. __________	9. __________	14. __________	19. __________	24. __________
5. __________	10. __________	15. __________	20. __________	25. __________

Introduction
class="introduction"

When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Note:
Chapter Objectives
By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: stemplots, histograms, and box plots.
- Recognize, describe, and calculate the measures of location of data: quartiles and percentiles.
- Recognize, describe, and calculate the measures of the center of data: mean, median, and mode.
- Recognize, describe, and calculate the measures of the spread of data: variance, standard deviation, and range.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study numerical and graphical ways to describe and display your data. This area of statistics is called **"Descriptive Statistics."** You will learn how to calculate, and even more importantly, how to interpret these measurements and graphs.

A statistical graph is a tool that helps you learn about the shape or distribution of a sample or a population. A graph can be a more effective way of presenting data than a mass of numbers because we can see where data clusters and where there are only a few data values. Newspapers and the Internet use graphs to show trends and to enable readers to compare facts and figures quickly. Statisticians often graph data first to get a picture of the data. Then, more formal tools may be applied.

Some of the types of graphs that are used to summarize and organize data are the dot plot, the bar graph, the histogram, the stem-and-leaf plot, the frequency polygon (a type of broken line graph), the pie chart, and the box plot. In this chapter, we will briefly look at stem-and-leaf plots, line graphs, and bar graphs, as well as frequency polygons, and time series graphs. Our emphasis will be on histograms and box plots.

Note:
NOTE
This book contains instructions for constructing a histogram and a box plot for the TI-83+ and TI-84 calculators. The Texas Instruments (TI) website provides additional instructions for using these calculators.

Stem-and-Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

One simple graph, the **stem-and-leaf graph** or **stemplot**, comes from the field of exploratory data analysis. It is a good choice when the data sets are small. To create the plot, divide each observation of data into a stem and a leaf. The leaf consists of a **final significant digit**. For example, 23 has stem two and leaf three. The number 432 has stem 43 and leaf two. Likewise, the number 5,432 has stem 543 and leaf two. The decimal 9.3 has stem nine and leaf three. Write the stems in a vertical line from smallest to largest. Draw a vertical line to the right of the stems. Then write the leaves in increasing order next to their corresponding stem.

Example:
For Susan Dean's spring pre-calculus class, scores for the first exam were as follows (smallest to largest):
33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Stem	**Leaf**
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

Stem-and-Leaf Graph

The stemplot shows that most scores fell in the 60s, 70s, 80s, and 90s. Eight out of the 31 scores or approximately 26% $\left(\frac{8}{31}\right)$ were in the 90s or 100, a fairly high number of As.

Note:

Try It
Exercise:

Problem:

For the Park City basketball team, scores for the last 30 games were as follows (smallest to largest):
32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61
Construct a stem plot for the data.

Solution:

Stem	**Leaf**
3	2 2 3 4 8
4	0 2 2 3 4 6 7 7 8 8 8 9
5	0 0 1 2 2 2 3 4 6 7 7
6	0 1

The stemplot is a quick way to graph data and gives an exact picture of the data. You want to look for an overall pattern and any outliers. An **outlier** is an observation of data that does not fit the rest of the data. It is sometimes called an **extreme value.** When you graph an outlier, it will appear not to fit the pattern of the graph. Some outliers are due to mistakes (for example, writing down 50 instead of 500) while others may indicate that something unusual is happening. It takes some background information to explain outliers, so we will cover them in more detail later.

Example:
The data are the distances (in kilometers) from a home to local supermarkets. Create a stemplot using the data:
1.1; 1.5; 2.3; 2.5; 2.7; 3.2; 3.3; 3.3; 3.5; 3.8; 4.0; 4.2; 4.5; 4.5; 4.7; 4.8; 5.5; 5.6; 6.5; 6.7; 12.3
Exercise:

Problem: Do the data seem to have any concentration of values?

Note:
NOTE
The leaves are to the right of the decimal.

Solution:

The value 12.3 may be an outlier. Values appear to concentrate at three and four kilometers.

Stem	Leaf
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Note:
Try It
Exercise:

Problem:

The following data show the distances (in miles) from the homes of off-campus statistics students to the college. Create a stem plot using the data and identify any outliers:

0.5; 0.7; 1.1; 1.2; 1.2; 1.3; 1.3; 1.5; 1.5; 1.7; 1.7; 1.8; 1.9; 2.0; 2.2; 2.5; 2.6; 2.8; 2.8; 2.8; 3.5; 3.8; 4.4; 4.8; 4.9; 5.2; 5.5; 5.7; 5.8; 8.0

Solution:

Stem	Leaf
0	5 7
1	1 2 2 3 3 5 5 7 7 8 9
2	0 2 5 6 8 8 8
3	5 8
4	4 8 9
5	2 5 7 8
6	
7	
8	0

The value 8.0 may be an outlier. Values appear to concentrate at one and two miles.

Example:
Exercise:

Problem:

A **side-by-side stem-and-leaf plot** allows a comparison of the two data sets in two columns. In a side-by-side stem-and-leaf plot, two sets of leaves share the same stem. The leaves are to the left and the right of the stems. [link] and [link] show the ages of presidents at their inauguration and at their death. Construct a side-by-side stem-and-leaf plot using this data.

Solution:

Ages at Inauguration		Ages at Death
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 4 2 2 1 1 1 1 1 0	5	3 6 6 7 7 8
9 8 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 0 1 1 1 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52

President	Age	President	Age	President	Age
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Presidential Ages at Inauguration

President	Age	President	Age	President	Age
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		

President	Age	President	Age	President	Age
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

Presidential Age at Death

Note:
Exercise:

Problem:

The table shows the number of wins and losses the Atlanta Hawks have had in 42 seasons. Create a side-by-side stem-and-leaf plot of these wins and losses.

Losses	Wins	Year	Losses	Wins	Year
34	48	1968–1969	41	41	1989–1990
34	48	1969–1970	39	43	1990–1991
46	36	1970–1971	44	38	1991–1992
46	36	1971–1972	39	43	1992–1993
36	46	1972–1973	25	57	1993–1994
47	35	1973–1974	40	42	1994–1995
51	31	1974–1975	36	46	1995–1996
53	29	1975–1976	26	56	1996–1997
51	31	1976–1977	32	50	1997–1998
41	41	1977–1978	19	31	1998–1999
36	46	1978–1979	54	28	1999–2000

Losses	Wins	Year	Losses	Wins	Year
32	50	1979–1980	57	25	2000–2001
51	31	1980–1981	49	33	2001–2002
40	42	1981–1982	47	35	2002–2003
39	43	1982–1983	54	28	2003–2004
42	40	1983–1984	69	13	2004–2005
48	34	1984–1985	56	26	2005–2006
32	50	1985–1986	52	30	2006–2007
25	57	1986–1987	45	37	2007–2008
32	50	1987–1988	35	47	2008–2009
30	52	1988–1989	29	53	2009–2010

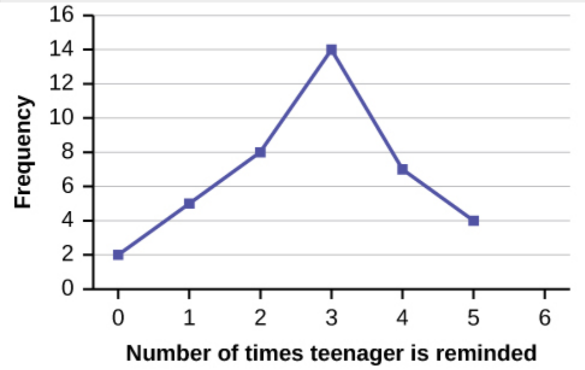
Solution:

Atlanta Hawks Wins and Losses		
Number of Wins		**Number of Losses**
3	1	9
9 8 8 6 5	2	5 5 9
8 7 6 6 5 5 4 3 1 1 1 1 0	3	0 2 2 2 2 4 4 5 6 6 6 9 9 9
8 8 7 6 6 6 3 3 3 2 2 1 1 0	4	0 0 1 1 2 4 5 6 6 7 7 8 9
7 7 6 3 2 0 0 0 0	5	1 1 1 2 3 4 4 6 7
	6	9

Another type of graph that is useful for specific data values is a **line graph**. In the particular line graph shown in [link], the ***x*-axis** (horizontal axis) consists of **data values** and the ***y*-axis** (vertical axis) consists of **frequency points**. The frequency points are connected using line segments.

Example:
In a survey, 40 mothers were asked how many times per week a teenager must be reminded to do his or her chores. The results are shown in [link] and in [link].

Number of times teenager is reminded	Frequency
0	2
1	5
2	8
3	14
4	7
5	4



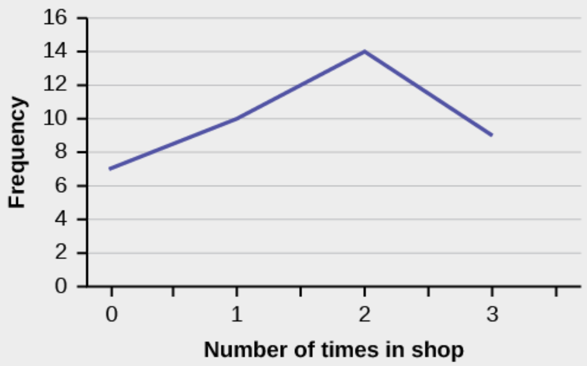
Note:
Try It
Exercise:

Problem:

In a survey, 40 people were asked how many times per year they had their car in the shop for repairs. The results are shown in [link]. Construct a line graph.

Number of times in shop	Frequency
0	7
1	10
2	14
3	9

Solution:



Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in [link] has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

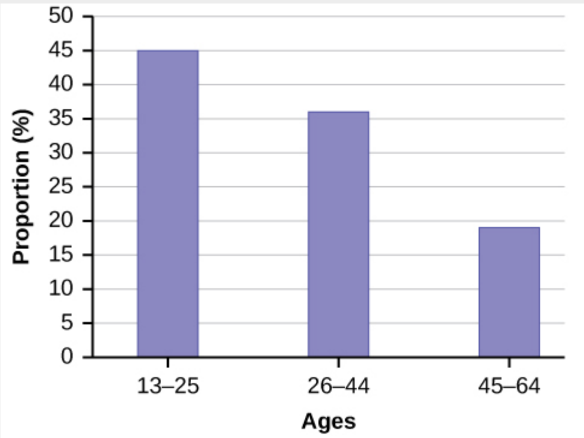
Example:
Exercise:

Problem:

By the end of 2011, Facebook had over 146 million users in the United States. [link] shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%

Solution:



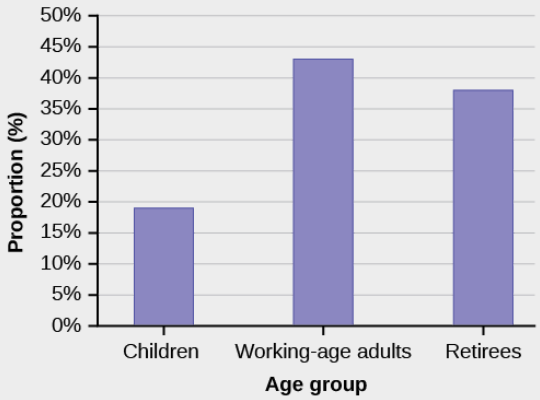
Note:
Try It
Exercise:

Problem:

The population in Park City is made up of children, working-age adults, and retirees. [link] shows the three age groups, the number of people in the town from each age group, and the proportion (%) of people in each age group. Construct a bar graph showing the proportions.

Age groups	Number of people	Proportion of population
Children	67,059	19%
Working-age adults	152,198	43%
Retirees	131,662	38%

Solution:



Example:
Exercise:

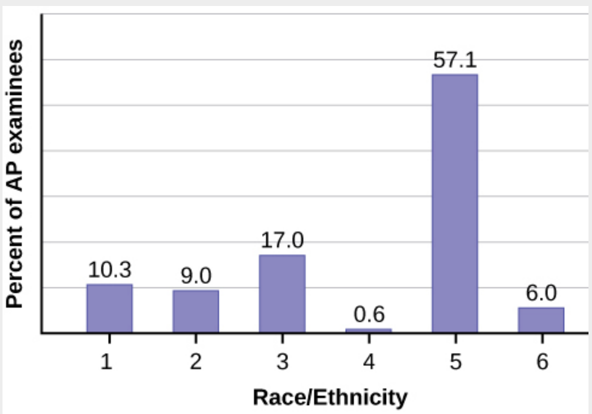
Problem:

The columns in [link] contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar graph with the student race or ethnicity (qualitative data) on the *x*-axis, and the Advanced Placement examinee population percentages on the *y*-axis.

Race/Ethnicity	AP Examinee Population	Overall Student Population
1 = Asian, Asian American or Pacific Islander	10.3%	5.7%
2 = Black or African American	9.0%	14.7%

Race/Ethnicity	AP Examinee Population	Overall Student Population
3 = Hispanic or Latino	17.0%	17.6%
4 = American Indian or Alaska Native	0.6%	1.1%
5 = White	57.1%	59.2%
6 = Not reported/other	6.0%	1.7%

Solution:



Note:
Try It
Exercise:

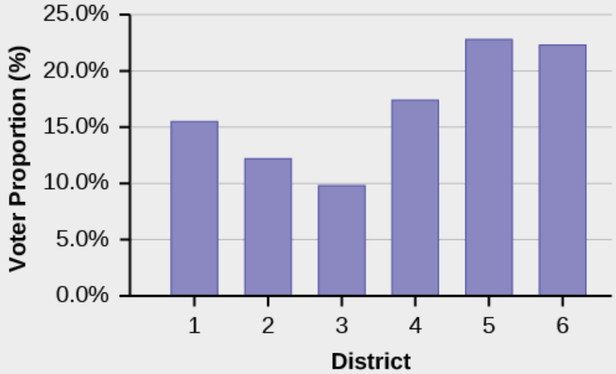
Problem:

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%

District	Registered voter population	Overall city population
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

Solution:



References

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics*, 2011. Available online at http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

"9th Annual AP Report to the Nation." CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goals-and-findings/promoting-equity (accessed September 13, 2013).

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Chapter Review

A **stem-and-leaf plot** is a way to plot data and look at the distribution. In a stem-and-leaf plot, all data values within a class are visible. The advantage in a stem-and-leaf plot is that all values are listed, unlike a histogram, which gives classes of data values. A **line graph** is often used to represent a set of data values in which a quantity varies with time. These graphs are useful for

finding trends. That is, finding a general pattern in data sets including temperature, sales, employment, company profit or cost over a period of time. A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used.

For each of the following data sets, create a stem plot and identify any outliers.

Exercise:

Problem:

The miles per gallon rating for 30 cars are shown below (lowest to highest).
19, 19, 19, 20, 21, 21, 25, 25, 25, 26, 26, 28, 29, 31, 31, 32, 32, 33, 34, 35, 36, 37, 37, 38, 38, 38, 38, 41, 43, 43

Solution:

Stem	**Leaf**
1	9 9 9
2	0 1 1 5 5 5 6 6 8 9
3	1 1 2 2 3 4 5 6 7 7 8 8 8 8
4	1 3 3

Exercise:

Problem:

The height in feet of 25 trees is shown below (lowest to highest).
25, 27, 33, 34, 34, 34, 35, 37, 37, 38, 39, 39, 39, 40, 41, 45, 46, 47, 49, 50, 50, 53, 53, 54, 54

Exercise:

Problem:

The data are the prices of different laptops at an electronics store. Round each value to the nearest ten.
249, 249, 260, 265, 265, 280, 299, 299, 309, 319, 325, 326, 350, 350, 350, 365, 369, 389, 409, 459, 489, 559, 569, 570, 610

Solution:

Stem	Leaf
2	5 5 6 7 7 8
3	0 0 1 2 3 3 5 5 5 7 7 9
4	1 6 9
5	6 7 7
6	1

Exercise:

Problem:

The data are daily high temperatures in a town for one month.
61, 61, 62, 64, 66, 67, 67, 67, 68, 69, 70, 70, 70, 71, 71, 72, 74, 74, 74, 75, 75, 75, 76, 76, 77, 78, 78, 79, 79, 95

For the next three exercises, use the data to construct a line graph.

Exercise:

Problem:

In a survey, 40 people were asked how many times they visited a store before making a major purchase. The results are shown in [link].

Number of times in store	Frequency
1	4
2	10
3	16

Number of times in store	Frequency
4	6
5	4

Solution:

Exercise:

Problem:

In a survey, several people were asked how many years it has been since they purchased a mattress. The results are shown in [link].

Years since last purchase	Frequency
0	2
1	8
2	13
3	22
4	16
5	9

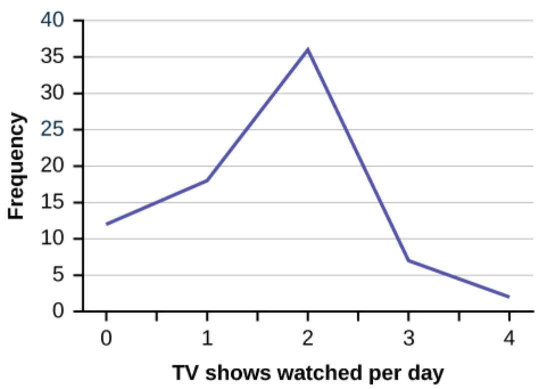
Exercise:

Problem:

Several children were asked how many TV shows they watch each day. The results of the survey are shown in [link].

Number of TV Shows	Frequency
0	12
1	18
2	36
3	7
4	2

Solution:



Exercise:

Problem:

The students in Ms. Ramirez's math class have birthdays in each of the four seasons. [link] shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the number of students.

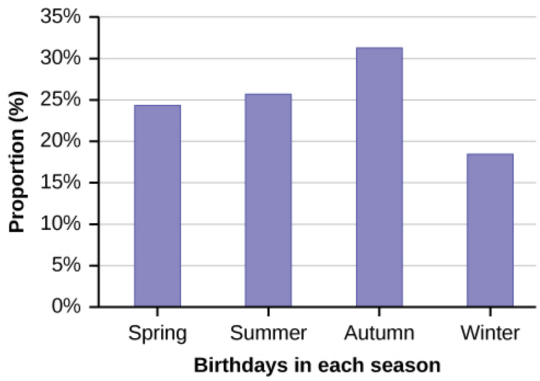
Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%

Exercise:

Problem:

Using the data from Mrs. Ramirez’s math class supplied in [link], construct a bar graph showing the percentages.

Solution:



Exercise:

Problem:

David County has six high schools. Each school sent students to participate in a county-wide science competition. [link] shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the population percentage of competitors from each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%

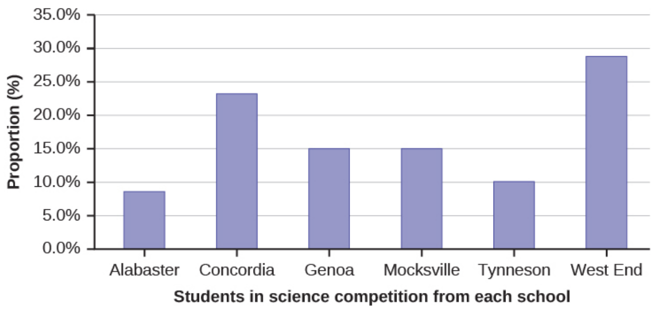
High School	Science competition population	Overall student population
Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%

Exercise:

Problem:

Use the data from the David County science competition supplied in [link]. Construct a bar graph that shows the county-wide population percentage of students at each school.

Solution:



Homework

Exercise:

Problem: Student grades on a chemistry exam were: 77, 78, 76, 81, 86, 51, 79, 82, 84, 99

a. Construct a stem-and-leaf plot of the data.
b. Are there any potential outliers? If so, which scores are they? Why do you consider them outliers?

Exercise:

Problem: [link] contains the 2010 obesity rates in U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

a. Use a random number generator to randomly pick eight states. Construct a bar graph of the obesity rates of those eight states.
b. Construct a bar graph for all the states beginning with the letter "A."

c. Construct a bar graph for all the states beginning with the letter "M."

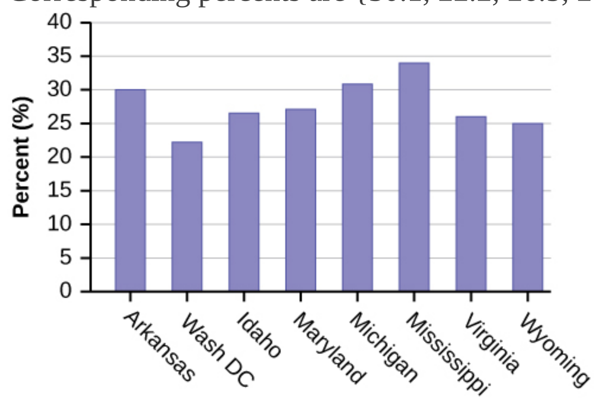
Solution:

a. Example solution for using the random number generator for the TI-84+ to generate a simple random sample of 8 states. Instructions are as follows.

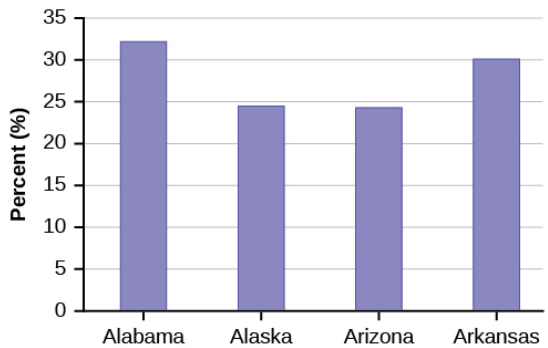
- Number the entries in the table 1–51 (Includes Washington, DC; Numbered vertically)
- Press MATH
- Arrow over to PRB
- Press 5:randInt(
- Enter 51,1,8)

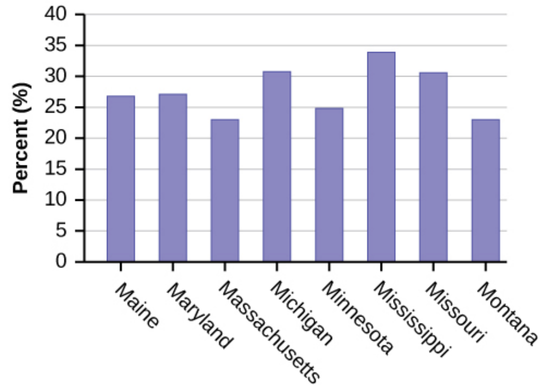
Eight numbers are generated (use the right arrow key to scroll through the numbers). The numbers correspond to the numbered states (for this example: {47 21 9 23 51 13 25 4}. If any numbers are repeated, generate a different number by using 5:randInt(51,1)). Here, the states (and Washington DC) are {Arkansas, Washington DC, Idaho, Maryland, Michigan, Mississippi, Virginia, Wyoming}.

Corresponding percents are {30.1, 22.2, 26.5, 27.1, 30.9, 34.0, 26.0, 25.1}.



b.




40
35
30
25
20
15
10
5
0
Percent (%)
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana

c.

Histograms, Frequency Polygons, and Time Series Graphs

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. A rule of thumb is to use a histogram when the data set consists of 100 values or more.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample. (Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:
Equation:

$$\text{RF} = \frac{f}{n}$$

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, f = 3, n = 40, and $RF = \frac{f}{n} = \frac{3}{40}$ = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

Example:
The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.
60; 60.5; 61; 61; 61.5
63.5; 63.5; 63.5
64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5
66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5
68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.
60 – 0.05 = 59.95 which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.
The largest value is 74, so 74 + 0.05 = 74.05 is the ending value.
Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

Equation:

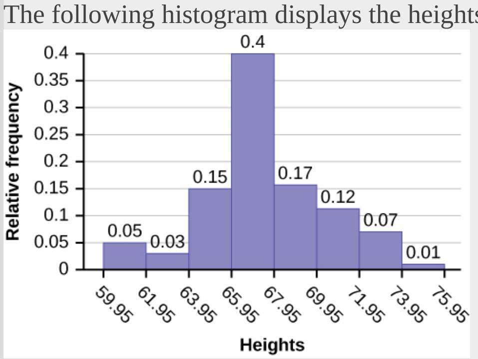
$$\frac{74.05 - 59.95}{8} = 1.76$$

Note:
NOTE
We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the number of bars or class intervals is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through 71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.
The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.



Note:
Try It
Exercise:

Problem:

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.
9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5; 12.5; 14

Solution:

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

$\frac{14.05-8.95}{6} = 0.85$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Example:
Create a histogram for the following data: the number of books bought by 50 part-time college students at ABC College.the number of books bought by 50 part-time college students at ABC College. The number of books is **discrete data**, since books are counted.
1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4
5; 5; 5; 5; 5
6; 6
Eleven students buy one book. Ten students buy two books. Sixteen students buy three books. Six students buy four books. Five students buy five books. Two students buy six books.
Because the data are integers, subtract 0.5 from 1, the smallest data value and add 0.5 to 6, the largest data value. Then the starting point is 0.5 and the ending value is 6.5.
Exercise:

Problem:

Next, calculate the width of each bar or class interval. If the data are discrete and there are not too many different values, a width that places the data values in the middle of the bar or class interval is the most convenient. Since the data consist of the numbers 1, 2, 3, 4, 5, 6, and the starting point is 0.5, a width of one places the 1 in the middle of the interval from 0.5 to 1.5, the 2 in the middle of the interval from 1.5 to 2.5, the 3 in the middle of the interval from 2.5 to 3.5, the 4 in the middle of the interval from _______ to _______, the 5 in the middle of the interval from _______ to _______, and the _______ in the middle of the interval from _______ to _______ .

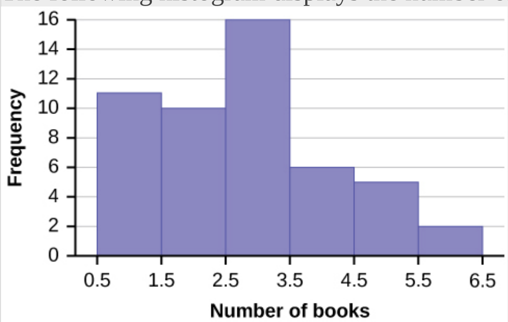
Solution:

- 3.5 to 4.5
- 4.5 to 5.5
- 6
- 5.5 to 6.5

Calculate the number of bars as follows:
Equation:

$$\frac{6.5 - 0.5}{\text{number of bars}} = 1$$

where 1 is the width of a bar. Therefore, bars = 6.
The following histogram displays the number of books on the *x*-axis and the frequency on the *y*-axis.



Note:
Go to [link]. There are calculator instructions for entering data and for creating a customized histogram. Create the histogram for [link].

- Press Y=. Press CLEAR to delete any equations.
- Press STAT 1:EDIT. If L1 has data in it, arrow up into the name L1, press CLEAR and then arrow down. If necessary, do the same for L2.
- Into L1, enter 1, 2, 3, 4, 5, 6.
- Into L2, enter 11, 10, 16, 6, 5, 2.
- Press WINDOW. Set Xmin = .5, Xmax = 6.5, Xscl = (6.5 – .5)/6, Ymin = –1, Ymax = 20, Yscl = 1, Xres = 1.
- Press 2nd Y=. Start by pressing 4:Plotsoff ENTER.
- Press 2nd Y=. Press 1:Plot1. Press ENTER. Arrow down to TYPE. Arrow to the 3rd picture (histogram). Press ENTER.
- Arrow down to Xlist: Enter L1 (2nd 1). Arrow down to Freq. Enter L2 (2nd 2).
- Press GRAPH.
- Use the TRACE key and the arrow keys to examine the histogram.

Note:
Try It
Exercise:

Problem:

The following data are the number of sports played by 50 student athletes. The number of sports is discrete data since sports are counted.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1
2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2

3; 3; 3; 3; 3; 3; 3; 3
20 student athletes play one sport. 22 student athletes play two sports. Eight student athletes play three sports.

Fill in the blanks for the following sentence. Since the data consist of the numbers 1, 2, 3, and the starting point is 0.5, a width of one places the 1 in the middle of the interval 0.5 to _____, the 2 in the middle of the interval from _____ to _____, and the 3 in the middle of the interval from _____ to _____.

Solution:

1.5
1.5 to 2.5
2.5 to 3.5

Example:
Exercise:

Problem: Using this data set, construct a histogram.

Number of Hours My Classmates Spent Playing Video Games on Weekends				
9.95	10	2.25	16.75	0
19.5	22.5	7.5	15	12.75
5.5	11	10	20.75	17.5
23	21.9	24	23.75	18
20	15	22.9	18.8	20.5

Solution:

Some values in this data set fall on boundaries for the class intervals. A value is counted in a class interval if it falls on the left boundary, but not if it falls on the right boundary. Different researchers may set up histograms for the same data in different ways. There is more than one correct way to set up a histogram.

Note:
Try It
Exercise:

Problem:

The following data represent the number of employees at various restaurants in New York City. Using this data, create a histogram.

22351526 40281820 25343942 24221927 22344020 38and 28
Use 10–19 as the first interval.

Note:
Count the money (bills and change) in your pocket or purse. Your instructor will record the amounts. As a class, construct a histogram displaying the data. Discuss how many intervals you think is appropriate. You may want to experiment with the number of intervals.

Frequency Polygons

Frequency polygons are analogous to line graphs, and just as line graphs make continuous data visually easy to interpret, so too do frequency polygons.

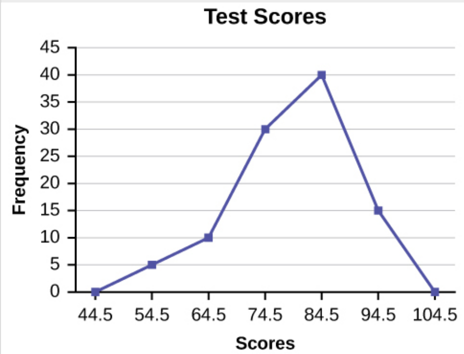
To construct a frequency polygon, first examine the data and decide on the number of intervals, or class intervals, to use on the *x*-axis and *y*-axis. After choosing the appropriate ranges, begin plotting the data points. After all the points are plotted, draw line segments to connect them.

Example:
A frequency polygon was constructed from the frequency table below.

Frequency Distribution for Calculus Final Test Scores

Lower Bound	Upper Bound	Frequency	Cumulative Frequency
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
79.5	89.5	40	85
89.5	99.5	15	100



The first label on the *x*-axis is 44.5. This represents an interval extending from 39.5 to 49.5. Since the lowest test score is 54.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 54.5 represents the next interval, or the first "real" interval from the table, and contains five scores. This reasoning is followed for each of the remaining intervals with the point 104.5 representing the interval from 99.5 to 109.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

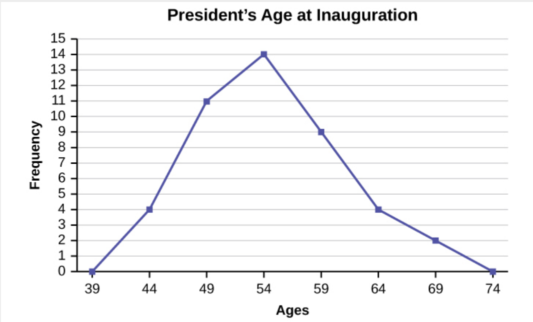
Note:
Try It
Exercise:

Problem: Construct a frequency polygon of U.S. Presidents' ages at inauguration shown in [link].

Age at Inauguration	Frequency
41.5–46.5	4
46.5–51.5	11
51.5–56.5	14
56.5–61.5	9
61.5–66.5	4
66.5–71.5	2

Solution:

The first label on the *x*-axis is 39. This represents an interval extending from 36.5 to 41.5. Since there are no ages less than 41.5, this interval is used only to allow the graph to touch the *x*-axis. The point labeled 44 represents the next interval, or the first “real” interval from the table, and contains four scores. This reasoning is followed for each of the remaining intervals with the point 74 representing the interval from 71.5 to 76.5. Again, this interval contains no data and is only used so that the graph will touch the *x*-axis. Looking at the graph, we say that this distribution is skewed because one side of the graph does not mirror the other side.

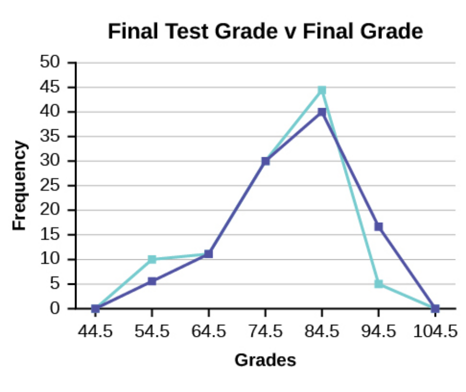


Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets.

Example:
We will construct an overlay frequency polygon comparing the scores from [link] with the students’ final numeric grade.

Frequency Distribution for Calculus Final Test Scores			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	5	5
59.5	69.5	10	15
69.5	79.5	30	45
79.5	89.5	40	85
89.5	99.5	15	100

Frequency Distribution for Calculus Final Grades			
Lower Bound	**Upper Bound**	**Frequency**	**Cumulative Frequency**
49.5	59.5	10	10
59.5	69.5	10	20
69.5	79.5	30	50
79.5	89.5	45	95
89.5	99.5	5	100



Suppose that we want to study the temperature range of a region for an entire month. Every day at noon we note the temperature and write this down in a log. A variety of statistical studies could be done with this data. We could find the mean or the median temperature for the month. We could construct a histogram displaying the number of days that temperatures reach a certain range of values. However, all of these methods ignore a portion of the data that we have collected.

One feature of the data that we may want to consider is that of time. Since each date is paired with the temperature reading for the day, we don't have to think of the data as being random. We can instead use the times given to impose a chronological order on the data. A graph that recognizes this ordering and displays the changing temperature as the month progresses is called a time series graph.

Constructing a Time Series Graph

To construct a time series graph, we must look at both pieces of our **paired data set**. We start with a standard Cartesian coordinate system. The horizontal axis is used to plot the date or time increments, and the vertical axis is used to plot the values of the variable that we are measuring. By doing this, we make each point on the graph correspond to a date and a measured quantity. The points on the graph are typically connected by straight lines in the order in which they occur.

Example:
Exercise:

Problem:

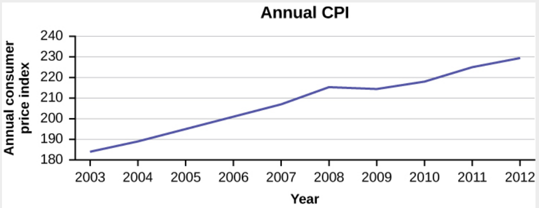
The following data shows the Annual Consumer Price Index, each month, for ten years. Construct a time series graph for the Annual Consumer Price Index data only.

Year	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**
2003	181.7	183.1	184.2	183.8	183.5	183.7	183.9
2004	185.2	186.2	187.4	188.0	189.1	189.7	189.4
2005	190.7	191.8	193.3	194.6	194.4	194.5	195.4
2006	198.3	198.7	199.8	201.5	202.5	202.9	203.5
2007	202.416	203.499	205.352	206.686	207.949	208.352	208.299
2008	211.080	211.693	213.528	214.823	216.632	218.815	219.964
2009	211.143	212.193	212.709	213.240	213.856	215.693	215.351
2010	216.687	216.741	217.631	218.009	218.178	217.965	218.011
2011	220.223	221.309	223.467	224.906	225.964	225.722	225.922
2012	226.665	227.663	229.392	230.085	229.815	229.478	229.104

Year	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Annual**
2003	184.6	185.2	185.0	184.5	184.3	184.0
2004	189.5	189.9	190.9	191.0	190.3	188.9
2005	196.4	198.8	199.2	197.6	196.8	195.3
2006	203.9	202.9	201.8	201.5	201.8	201.6
2007	207.917	208.490	208.936	210.177	210.036	207.342
2008	219.086	218.783	216.573	212.425	210.228	215.303
2009	215.834	215.969	216.177	216.330	215.949	214.537
2010	218.312	218.439	218.711	218.803	219.179	218.056

Year	Aug	Sep	Oct	Nov	Dec	Annual
2011	226.545	226.889	226.421	226.230	225.672	224.939
2012	230.379	231.407	231.317	230.221	229.601	229.594

Solution:



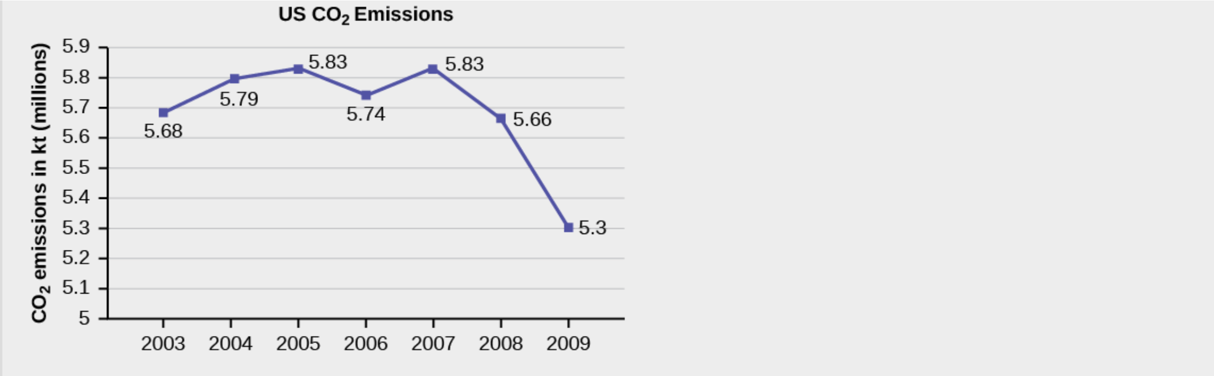
Note:
Try It
Exercise:

Problem:

The following table is a portion of a data set from www.worldbank.org. Use the table to construct a time series graph for CO_2 emissions for the United States.

CO2 Emissions			
	Ukraine	**United Kingdom**	**United States**
2003	352,259	540,640	5,681,664
2004	343,121	540,409	5,790,761
2005	339,029	541,990	5,826,394
2006	327,797	542,045	5,737,615
2007	328,357	528,631	5,828,697
2008	323,657	522,247	5,656,839
2009	272,176	474,579	5,299,563

Solution:



Uses of a Time Series Graph

Time series graphs are important tools in various applications of statistics. When recording values of the same variable over an extended period of time, sometimes it is difficult to discern any trend or pattern. However, once the same data points are displayed graphically, some features jump out. Time series graphs make trends easy to spot.

References

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Chapter Review

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets. A frequency polygon can also be used when graphing large data sets with data points that repeat. The data usually goes on *y*-axis with the frequency being graphed on the *x*-axis. Time series graphs can be helpful when looking at large amounts of data for one variable over a period of time.

Exercise:

Problem:

Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Complete the table.

Data Value (# cars)	Frequency	Relative Frequency	Cumulative Relative Frequency

Exercise:

Problem: What does the frequency column in [link] sum to? Why?

Solution:

65

Exercise:

Problem: What does the relative frequency column in [link] sum to? Why?

Exercise:

Problem: What is the difference between relative frequency and frequency for each data value in [link]?

Solution:

The relative frequency shows the *proportion* of data points that have each value. The frequency tells the *number* of data points that have each value.

Exercise:

Problem:

What is the difference between cumulative relative frequency and relative frequency for each data value?

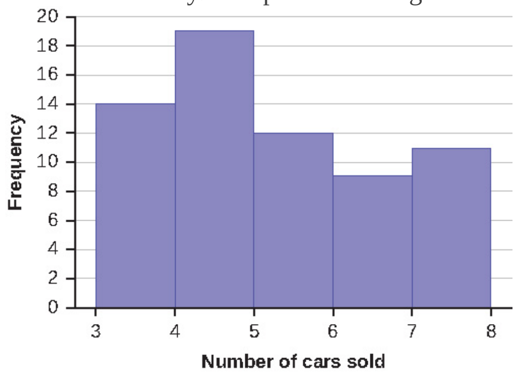
Exercise:

Problem:

To construct the histogram for the data in [link], determine appropriate minimum and maximum *x* and *y* values and the scaling. Sketch the histogram. Label the horizontal and vertical axes with words. Include numerical scaling.

Solution:

Answers will vary. One possible histogram is shown:



Exercise:

Problem: Construct a frequency polygon for the following:

a.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

b.

Actual Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

c.

Tar (mg) in Nonfiltered Cigarettes	Frequency
10–13	1
14–17	0
18–21	15
22–25	7
26–29	2

Exercise:

Problem:

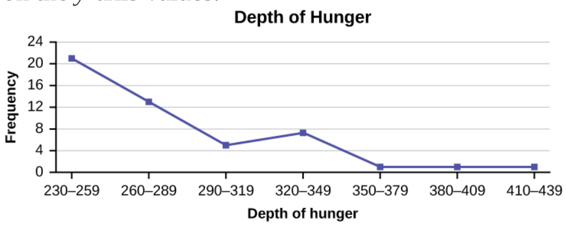
Construct a frequency polygon from the frequency distribution for the 50 highest ranked countries for depth of hunger.

Depth of Hunger	Frequency
230–259	21
260–289	13
290–319	5
320–349	7
350–379	1
380–409	1

Depth of Hunger	Frequency
410–439	1

Solution:

Find the midpoint for each class. These will be graphed on the *x*-axis. The frequency values will be graphed on the *y*-axis values.



Exercise:

Problem:

Use the two frequency tables to compare the life expectancy of men and women from 20 randomly selected countries. Include an overlayed frequency polygon and discuss the shapes of the distributions, the center, the spread, and any outliers. What can we conclude about the life expectancy of women compared to men?

Life Expectancy at Birth – Women	Frequency
49–55	3
56–62	3
63–69	1
70–76	3
77–83	8
84–90	2

Life Expectancy at Birth – Men	Frequency
49–55	3
56–62	3

Life Expectancy at Birth – Men	**Frequency**
63–69	1
70–76	1
77–83	7
84–90	5

Exercise:

Problem:

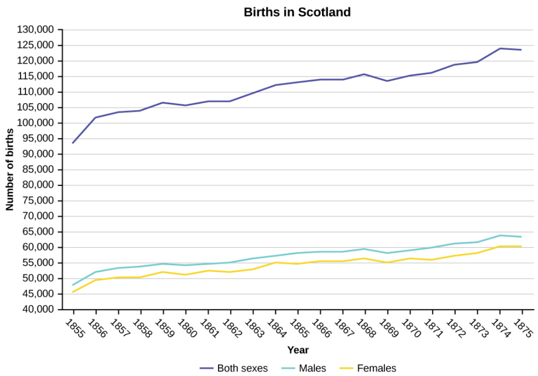
Construct a times series graph for (a) the number of male births, (b) the number of female births, and (c) the total number of births.

Sex/Year	1855	1856	1857	1858	1859	1860	1861
Female	45,545	49,582	50,257	50,324	51,915	51,220	52,403
Male	47,804	52,239	53,158	53,694	54,628	54,409	54,606
Total	93,349	101,821	103,415	104,018	106,543	105,629	107,009

Sex/Year	1862	1863	1864	1865	1866	1867	1868	18
Female	51,812	53,115	54,959	54,850	55,307	55,527	56,292	55
Male	55,257	56,226	57,374	58,220	58,360	58,517	59,222	58
Total	107,069	109,341	112,333	113,070	113,667	114,044	115,514	11

Sex/Year	1870	1871	1872	1873	1874	1875
Female	56,431	56,099	57,472	58,233	60,109	60,146
Male	58,959	60,029	61,293	61,467	63,602	63,432
Total	115,390	116,128	118,765	119,700	123,711	123,578

Solution:



Exercise:

Problem:

The following data sets list full time police per 100,000 citizens along with homicides per 100,000 citizens for the city of Detroit, Michigan during the period from 1961 to 1973.

Year	1961	1962	1963	1964	1965	1966	1967
Police	260.35	269.8	272.04	272.96	272.51	261.34	268.89
Homicides	8.6	8.9	8.52	8.89	13.07	14.57	21.36

Year	1968	1969	1970	1971	1972	1973
Police	295.99	319.87	341.43	356.59	376.69	390.19
Homicides	28.03	31.49	37.39	46.26	47.24	52.33

a. Construct a double time series graph using a common *x*-axis for both sets of data.
b. Which variable increased the fastest? Explain.
c. Did Detroit's increase in police officers have an impact on the murder rate? Explain.

Homework

Exercise:

Problem:

Suppose that three book publishers were interested in the number of fiction paperbacks adult consumers purchase per month. Each publisher conducted a survey. In the survey, adult consumers were asked the number of fiction paperbacks they had purchased the previous month. The results are as follows:

# of books	Freq.	Rel. Freq.
0	10	
1	12	
2	16	
3	12	
4	8	
5	6	
6	2	
8	2	

Publisher A

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

Publisher B

# of books	Freq.	Rel. Freq.
0–1	20	
2–3	35	
4–5	12	
6–7	2	
8–9	1	

Publisher C

a. Find the relative frequencies for each survey. Write them in the charts.
b. Using either a graphing calculator, computer, or by hand, use the frequency column to construct a histogram for each publisher's survey. For Publishers A and B, make bar widths of one. For Publisher C, make bar widths of two.
c. In complete sentences, give two reasons why the graphs for Publishers A and B are not identical.
d. Would you have expected the graph for Publisher C to look like the other two graphs? Why or why not?
e. Make new histograms for Publisher A and Publisher B. This time, make bar widths of two.
f. Now, compare the graph for Publisher C to the new graphs for Publishers A and B. Are the graphs more similar or more different? Explain your answer.

Exercise:

Problem:

Often, cruise ships conduct all on-board transactions, with the exception of gambling, on a cashless basis. At the end of the cruise, guests pay one bill that covers all onboard transactions. Suppose that 60 single travelers and 70 couples were surveyed as to their on-board bills for a seven-day cruise from Los Angeles to the Mexican Riviera. Following is a summary of the bills for each group.

Amount($)	Frequency	Rel. Frequency
51–100	5	
101–150	10	
151–200	15	
201–250	15	
251–300	10	
301–350	5	

Singles

Amount($)	Frequency	Rel. Frequency
100–150	5	
201–250	5	
251–300	5	
301–350	5	
351–400	10	
401–450	10	
451–500	10	
501–550	10	
551–600	5	
601–650	5	

Couples

a. Fill in the relative frequency for each group.
b. Construct a histogram for the singles group. Scale the x-axis by $50 widths. Use relative frequency on the y-axis.
c. Construct a histogram for the couples group. Scale the x-axis by $50 widths. Use relative frequency on the y-axis.
d. Compare the two graphs:
    i. List two similarities between the graphs.
    ii. List two differences between the graphs.
    iii. Overall, are the graphs more similar or different?
e. Construct a new graph for the couples by hand. Since each couple is paying for two individuals, instead of scaling the x-axis by $50, scale it by $100. Use relative frequency on the y-axis.
f. Compare the graph for the singles with the new graph for the couples:
    i. List two similarities between the graphs.
    ii. Overall, are the graphs more similar or different?
g. How did scaling the couples graph differently change the way you compared it to the singles graph?
h. Based on the graphs, do you think that individuals spend the same amount, more or less, as singles as they do person by person as a couple? Explain why in one or two complete sentences.

Solution:

Amount($)	Frequency	Relative Frequency

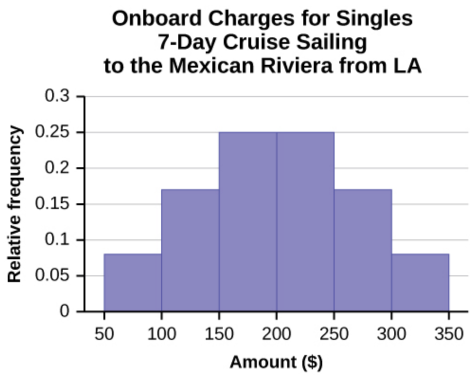
Amount($)	**Frequency**	**Relative Frequency**
51–100	5	0.08
101–150	10	0.17
151–200	15	0.25
201–250	15	0.25
251–300	10	0.17
301–350	5	0.08

Singles

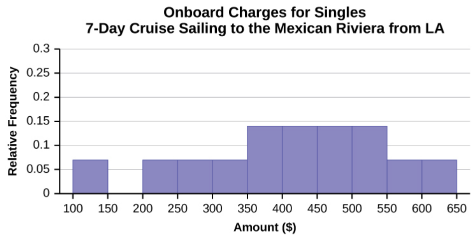
Amount($)	**Frequency**	**Relative Frequency**
100–150	5	0.07
201–250	5	0.07
251–300	5	0.07
301–350	5	0.07
351–400	10	0.14
401–450	10	0.14
451–500	10	0.14
501–550	10	0.14
551–600	5	0.07
601–650	5	0.07

Couples

a. See [link] and [link].
b. In the following histogram data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where both boundary values are included).



c. In the following histogram, the data values that fall on the right boundary are counted in the class interval, while values that fall on the left boundary are not counted (with the exception of the first interval where values on both boundaries are included).



d. Compare the two graphs:

   i. Answers may vary. Possible answers include:

      - Both graphs have a single peak.
      - Both graphs use class intervals with width equal to $50.

   ii. Answers may vary. Possible answers include:

      - The couples graph has a class interval with no values.
      - It takes almost twice as many class intervals to display the data for couples.

   iii. Answers may vary. Possible answers include: The graphs are more similar than different because the overall patterns for the graphs are the same.

e. Check student's solution.
f. Compare the graph for the Singles with the new graph for the Couples:

   i.
      - Both graphs have a single peak.
      - Both graphs display 6 class intervals.
      - Both graphs show the same general pattern.

   ii. Answers may vary. Possible answers include: Although the width of the class intervals for couples is double that of the class intervals for singles, the graphs are more similar than they are different.

g. Answers may vary. Possible answers include: You are able to compare the graphs interval by interval. It is easier to compare the overall patterns with the new scale on the Couples graph. Because a couple represents two individuals, the new scale leads to a more accurate comparison.
h. Answers may vary. Possible answers include: Based on the histograms, it seems that spending does not vary much from singles to individuals who are part of a couple. The overall patterns are the same. The range of spending for couples is approximately double the range for individuals.

Exercise:

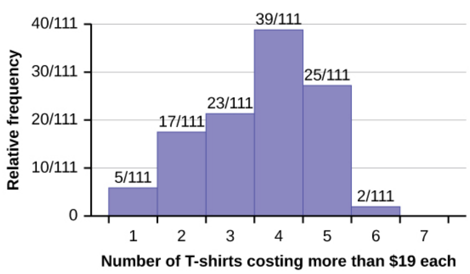
Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows.

# of movies	Frequency	Relative Frequency	Cumulative Relative Frequency
0	5		
1	9		
2	6		
3	4		
4	1		

a. Construct a histogram of the data.
b. Complete the columns of the chart.

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.



Exercise:

Problem:

The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

a. 21
b. 59
c. 41
d. Cannot be determined

Solution:

c

Exercise:

Problem:

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

a. cluster
b. simple random
c. stratified
d. convenience

Exercise:

Problem: Following are the 2010 obesity rates by U.S. states and Washington, DC.

State	Percent (%)	State	Percent (%)	State	Percent (%)
Alabama	32.2	Kentucky	31.3	North Dakota	27.2
Alaska	24.5	Louisiana	31.0	Ohio	29.2
Arizona	24.3	Maine	26.8	Oklahoma	30.4
Arkansas	30.1	Maryland	27.1	Oregon	26.8
California	24.0	Massachusetts	23.0	Pennsylvania	28.6
Colorado	21.0	Michigan	30.9	Rhode Island	25.5
Connecticut	22.5	Minnesota	24.8	South Carolina	31.5
Delaware	28.0	Mississippi	34.0	South Dakota	27.3
Washington, DC	22.2	Missouri	30.5	Tennessee	30.8
Florida	26.6	Montana	23.0	Texas	31.0
Georgia	29.6	Nebraska	26.9	Utah	22.5
Hawaii	22.7	Nevada	22.4	Vermont	23.2
Idaho	26.5	New Hampshire	25.0	Virginia	26.0
Illinois	28.2	New Jersey	23.8	Washington	25.5

State	Percent (%)	State	Percent (%)	State	Percent (%)
Indiana	29.6	New Mexico	25.1	West Virginia	32.5
Iowa	28.4	New York	23.9	Wisconsin	26.3
Kansas	29.4	North Carolina	27.8	Wyoming	25.1

Construct a bar graph of obesity rates of your state and the four states closest to your state. Hint: Label the *x*-axis with the states.

Solution:

Answers will vary.

Glossary

Frequency
: the number of times a value of the data occurs

Histogram
: a graphical representation in x-y form of the distribution of data in a data set; x represents the data and y represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency
: the ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes

Measures of the Location of the Data

The common measures of location are **quartiles** and **percentiles**

Quartiles are special percentiles. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, M, is called both the second quartile and the 50th percentile.

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.
1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1
Ordered from smallest to largest:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.
Equation:

$$\frac{6.8 + 7.2}{2} = 7$$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, Q_1, is the middle value of the lower half of the data, and the third quartile, Q_3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.
1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, $Q3$, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$IQR = Q_3 – Q_1$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(IQR) below the first quartile or more than (1.5)(IQR) above the third quartile**. Potential outliers always require further investigation.

Note:
NOTE
A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

Example:
Exercise:

Problem:

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars.
389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution:

Order the data from smallest to largest.
114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

M = 488,800

$Q_1 = \frac{230{,}500 + 387{,}000}{2} = 308{,}750$

$Q_3 = \frac{639{,}000 + 659{,}000}{2} = 649{,}000$

IQR = 649,000 – 308,750 = 340,250

(1.5)(IQR) = (1.5)(340,250) = 510,375

Q_1 – (1.5)(IQR) = 308,750 – 510,375 = –201,625

Q_3 + (1.5)(IQR) = 649,000 + 510,375 = 1,159,375

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Note:
Try It
Exercise:

Problem:

For the following 11 salaries, calculate the *IQR* and determine if any salaries are outliers. The salaries are in dollars.

\$33,000 \$64,500 \$28,000 \$54,000 \$72,000 \$68,500 \$69,000 \$42,000 \$54,000 \$120,000 \$40,500

Solution:

Order the data from smallest to largest.

\$28,000 \$33,000 \$40,500 \$42,000 \$54,000 \$54,000 \$64,500 \$68,500 \$69,000 \$72,000 \$120,000

Median = \$54,000

Q_1 = \$40,500

Q_3 = \$69,000

IQR = \$69,000 – \$40,500 = \$28,500

(1.5)(*IQR*) = (1.5)(\$28,500) = \$42,750

Q_1 – (1.5)(*IQR*) = \$40,500 – \$42,750 = –\$2,250

Q_3 + (1.5)(*IQR*) = \$69,000 + \$42,750 = \$111,750

No salary is less than –\$2,250. However, \$120,000 is more than \$11,750, so \$120,000 is a potential outlier.

Example:
Exercise:

Problem:

For the two data sets in the test scores example, find the following:

a. The interquartile range. Compare the two interquartile ranges.
b. Any outliers in either set.

Solution:

The five number summary for the day and night classes is

	Minimum	**Q_1**	**Median**	**Q_3**	**Maximum**
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

a. The IQR for the day group is $Q_3 – Q_1$ = 82.5 – 56 = 26.5

   The IQR for the night group is $Q_3 – Q_1$ = 89 – 78 = 11

   The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.
b. Day class outliers are found using the IQR times 1.5 rule. So,

   - Q_1 - *IQR*(1.5) = 56 – 26.5(1.5) = 16.25
   - Q_3 + *IQR*(1.5) = 82.5 + 26.5(1.5) = 122.25

   Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

   Night class outliers are calculated as:

- $Q_1 - IQR\,(1.5) = 78 - 11(1.5) = 61.5$
- $Q_3 + \text{IQR}(1.5) = 89 + 11(1.5) = 105.5$

For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Note:
Try It
Exercise:

Problem:

Find the interquartile range for the following two data sets and compare them.

Test Scores for Class *A*
69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94
Test Scores for Class *B*
90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Solution:

Class *A*

Order the data from smallest to largest.

65 66 67 69 69 76 77 77 79 80 81 83 85 89 90 91 94 96 98 99

$\text{Median} = \frac{80+81}{2} = 80.5$

$Q_1 = \frac{69+76}{2} = 72.5$

$Q_3 = \frac{90+91}{2} = 90.5$

$IQR = 90.5 - 72.5 = 18$

Class *B*

Order the data from smallest to largest.

68 68 70 71 72 73 75 78 79 80 80 90 90 92 92 95 95 97 99 100

$\text{Median} = \frac{80+80}{2} = 80$

$Q_1 = \frac{72+73}{2} = 72.5$

$Q_3 = \frac{92+95}{2} = 93.5$

$IQR = 93.5 - 72.5 = 21$

The data for Class *B* has a larger *IQR*, so the scores between Q_3 and Q_1 (middle 50%) for the data for Class *B* are more spread out and not clustered about the median.

Example:
Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were:

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
6	7	0.14	0.28
7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Find the 28th percentile. Notice the 0.28 in the "cumulative relative frequency" column. Twenty-eight percent of 50 data values is 14 values. There are 14 values less than the 28th percentile. They include the two 4s, the five 5s, and the seven 6s. The 28th percentile is between the last six and the first seven. **The 28th percentile is 6.5.**

Find the median. Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th percentile or the second quartile. 50% of 50 is 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th percentile is between the 25th, or seven, and 26th, or seven, values. **The median is seven.**

Find the third quartile. The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the fours, fives, sixes and sevens, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th percentile, then, must be an eight**. Another way to look at the problem is to find 75% of 50, which is 37.5, and round up to 38. The third quartile, Q_3, is the 38th value, which is an eight. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Note:
Try it
Exercise:

Problem:

Forty bus drivers were asked how many hours they spend each day running their routes (rounded to the nearest hour). Find the 65th percentile.

Amount of time spent on route (hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
2	12	0.30	0.30
3	14	0.35	0.65
4	10	0.25	0.90
5	4	0.10	1.00

Solution:

The 65th percentile is between the last three and the first four.

The 65th percentile is 3.5.

Example:
Exercise:

Problem: Using [link]:

a. Find the 80th percentile.
b. Find the 90th percentile.
c. Find the first quartile. What is another name for the first quartile?

Solution:

Using the data from the frequency table, we have:

a. The 80th percentile is between the last eight and the first nine in the table (between the 40th and 41st values). Therefore, we need to take the mean of the 40th an 41st values. The 80th percentile $= \frac{8+9}{2} = 8.5$
b. The 90th percentile will be the 45th data value (location is 0.90(50) = 45) and the 45th data value is nine.
c. Q_1 is also the 25th percentile. The 25th percentile location calculation: P_{25} = 0.25(50) = 12.5 ≈ 13 the 13th data value. Thus, the 25th percentile is six.

Note:
Try It
Exercise:

Problem:

Refer to the [link]. Find the third quartile. What is another name for the third quartile?

Solution:

The third quartile is the 75th percentile, which is four. The 65th percentile is between three and four, and the 90th percentile is between four and 5.75. The third quartile is between 65 and 90, so it must be four.

Note:
Collaborative Statistics

Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions:

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Construct two different histograms. For each, starting value = _____ ending value = ____.
4. Find the median, first quartile, and third quartile.
5. Construct a table of the data to find the following:

    a. the 10^{th} percentile
    b. the 70^{th} percentile
    c. the percent of students who own less than four sweaters

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

Example:
Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70^{th} percentile.
b. Find the 83^{rd} percentile.

Solution:

a.
- $k = 70$
- i = the index
- $n = 29$

$i = \frac{k}{100}(n + 1) = (\frac{70}{100})(29 + 1) = 21$. Twenty-one is an integer, and the data value in the 21^{st} position in the ordered data set is 64. The 70^{th} percentile is 64 years.

b.
- $k = 83^{rd}$ percentile
- i = the index
- $n = 29$

$i = \frac{k}{100}(n + 1) =)\frac{83}{100})(29 + 1) = 24.9$, which is NOT an integer. Round it down to 24 and up to 25. The age in the 24^{th} position is 71 and the age in the 25^{th} position is 72. Average 71 and 72. The 83^{rd} percentile is 71.5 years.

Note:
Try It
Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77
Calculate the 20^{th} percentile and the 55^{th} percentile.

Solution:

k = 20. Index = i = $\frac{k}{100}(n+1) = \frac{20}{100}(29 + 1)$ = 6. The age in the sixth position is 27. The 20^{th} percentile is 27 years.

k = 55. Index = i = $\frac{k}{100}(n+1) = \frac{55}{100}(29 + 1)$ = 16.5. Round down to 16 and up to 17. The age in the 16^{th} position is 52 and the age in the 17^{th} position is 55. The average of 52 and 55 is 53.5. The 55^{th} percentile is 53.5 years.

Note:
NOTE
You can calculate percentiles using calculators and computers. There are a variety of online calculators.

A Formula for Finding the Percentile of a Value in a Data Set

- Order the data from smallest to largest.
- x = the number of data values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile.
- y = the number of data values equal to the data value for which you want to find the percentile.
- n = the total number of data.
- Calculate $\frac{x+0.5y}{n}(100)$. Then round to the nearest integer.

Example:
Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*
18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile for 58.
b. Find the percentile for 25.

Solution:

a. Counting from the bottom of the list, there are 18 data values less than 58. There is one value of 58.

   x = 18 and y = 1. $\frac{x+0.5y}{n}(100) = \frac{18+0.5(1)}{29}(100) = 63.80$. 58 is the 64th percentile.
b. Counting from the bottom of the list, there are three data values less than 25. There is one value of 25.

   x = 3 and y = 1. $\frac{x+0.5y}{n}(100) = \frac{3+0.5(1)}{29}(100) = 12.07$. Twenty-five is the 12th percentile.

Note:
Try It
Exercise:

Problem:

Listed are 30 ages for Academy Award winning best actors in order from smallest to largest.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Find the percentiles for 47 and 31.

Solution:

Percentile for 47: Counting from the bottom of the list, there are 15 data values less than 47. There is one value of 47.

x = 15 and y = 1. $\frac{x+0.5y}{n}(100) = \frac{15+0.5(1)}{29}(100) = 53.45$. 47 is the 53rd percentile.

Percentile for 31: Counting from the bottom of the list, there are eight data values less than 31. There are two values of 31.

x = 15 and y = 2. $\frac{x+0.5y}{n}(100) = \frac{15+0.5(2)}{29}(100) = 31.03$. 31 is the 31st percentile.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Note:
NOTE
When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

Example:
Exercise:

Problem:

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

Solution:

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Note:
Try It
Exercise:

Problem:

For the 100-meter dash, the third quartile for times for finishing the race was 11.5 seconds. Interpret the third quartile in the context of the situation.

Solution:

Twenty-five percent of runners finished the race in 11.5 seconds or more. Seventy-five percent of runners finished the race in 11.5 seconds or less. A lower percentile is good because finishing a race more quickly is desirable.

Example:
Exercise:

Problem:

On a 20 question math test, the 70^{th} percentile for number of correct answers was 16. Interpret the 70^{th} percentile in the context of this situation.

Solution:

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Note:
Try It
Exercise:

Problem:

On a 60 point written assignment, the 80^{th} percentile for the number of points earned was 49. Interpret the 80^{th} percentile in the context of this situation.

Solution:

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

Example:
Exercise:

Problem:

At a community college, it was found that the 30^{th} percentile of credit units that students are enrolled for is seven units. Interpret the 30^{th} percentile in the context of this situation.

Solution:

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Note:
Try It
Exercise:

Problem:

During a season, the 40^{th} percentile for points scored per player in a game is eight. Interpret the 40^{th} percentile in the context of this situation.

Solution:

Forty percent of players scored eight points or fewer. Sixty percent of players scored eight points or more. A higher percentile is good because getting more points in a basketball game is desirable.

Example:
Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym. The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.
0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes
10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes;
30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes
Determine the following five values.

- Min = 0
- Q_1 = 20
- Med = 40
- Q_3 = 60
- Max = 300

If you were the principal, would you be justified in purchasing new fitness equipment? Since 75% of the students exercise for 60 minutes or less daily, and since the *IQR* is 40 minutes (60 – 20 = 40), we know that half of the students surveyed exercise between 20 minutes and 60 minutes daily. This seems a reasonable amount of time spent exercising, so the principal would be justified in purchasing the new equipment.
However, the principal needs to be careful. The value 300 appears to be a potential outlier.
$Q_3 + 1.5(IQR) = 60 + (1.5)(40) = 120$.
The value 300 is greater than 120 so it is a potential outlier. If we delete it and calculate the five values, we get the following values:

- Min = 0
- Q_1 = 20
- Q_3 = 60
- Max = 120

We still have 75% of the students exercising for 60 minutes or less daily and half of the students exercising between 20 and 60 minutes a day. However, 15 students is a small sample and the principal should survey more students to be sure of his survey results.

References

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Chapter Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles. Percentiles are used to compare and interpret data. For example, an observation at the 50^{th} percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25^{th} percentile,the second quartile (Q_2 or median) is 50^{th} percentile, and the third quartile (Q_3) is the the 75^{th} percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 - IQR(1.5)$

Formula Review

$i = \left(\frac{k}{100}\right)(n+1)$

where *i* = the ranking or position of a data value,

k = the kth percentile,

n = total number of data.

Expression for finding the percentile of a data value: $\left(\frac{x+0.5y}{n}\right)(100)$

where *x* = the number of values counting from the bottom of the data list up to but not including the data value for which you want to find the percentile,

y = the number of data values equal to the data value for which you want to find the percentile,

n = total number of data

Exercise:

Problem:

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 40th percentile.
b. Find the 78th percentile.

Solution:

a. The 40th percentile is 37 years.
b. The 78th percentile is 70 years.

Exercise:

Problem:

Listed are 32 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 18; 21; 22; 25; 26; 27; 29; 30; 31; 31; 33; 36; 37; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the percentile of 37.
b. Find the percentile of 72.

Exercise:

Problem:

Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking?

Solution:

Jesse graduated 37th out of a class of 180 students. There are 180 – 37 = 143 students ranked below Jesse. There is one rank of 37.

$x = 143$ and $y = 1$. $\frac{x+0.5y}{n}(100) = \frac{143+0.5(1)}{180}(100) = 79.72$. Jesse's rank of 37 puts him at the 80th percentile.

Exercise:

Problem:

a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
c. A bicyclist in the 90th percentile of a bicycle race completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise:

Problem:

a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Solution:

a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed which is faster.
b. 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Exercise:

Problem:

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise:

Problem:

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Solution:

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Exercise:

Problem:

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78^{th} percentile. Should Li be pleased or upset by this result? Explain.

Exercise:

Problem:

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90^{th} percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90^{th} percentile in the context of this problem.

Solution:

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Exercise:

Problem:

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
b. Students whose GPAs are at or above the 96^{th} percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percentage of students from each high school are "eligible in the local context"?

Exercise:

Problem:

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34^{th} percentile. The 34^{th} percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

Solution:

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

Use the following information to answer the next six exercises. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Exercise:

Problem: First quartile = ________

Exercise:

Problem: Second quartile = median = 50^{th} percentile = ________

Solution:

4

Exercise:

Problem: Third quartile = ________

Exercise:

Problem: Interquartile range (IQR) = _____ – _____ = _____

Solution:

6 – 4 = 2

Exercise:

Problem: 10th percentile = ________

Exercise:

Problem: 70th percentile = ________

Solution:

6

Homework

Exercise:

Problem:

The median age for U.S. blacks currently is 30.9 years; for U.S. whites it is 42.3 years.

a. Based upon this information, give two reasons why the black median age could be lower than the white median age.
b. Does the lower median age for blacks necessarily mean that blacks die younger than whites? Why or why not?
c. How might it be possible for blacks and whites to die at approximately the same age, but for the median age for whites to be higher?

Exercise:

Problem:

Six hundred adult Americans were asked by telephone poll, "What do you think constitutes a middle-class income?" The results are in [link]. Also, include left endpoint, but not the right endpoint.

Salary ($)	Relative Frequency
< 20,000	0.02
20,000–25,000	0.09
25,000–30,000	0.19
30,000–40,000	0.26
40,000–50,000	0.18
50,000–75,000	0.17
75,000–99,999	0.02
100,000+	0.01

a. What percentage of the survey answered "not sure"?
b. What percentage think that middle-class is from $25,000 to $50,000?
c. Construct a histogram of the data.

   i. Should all bars have the same width, based on the data? Why or why not?
   ii. How should the <20,000 and the 100,000+ intervals be handled? Why?

d. Find the 40^{th} and 80^{th} percentiles
e. Construct a bar graph of the data

Solution:

a. 1 – (0.02+0.09+0.19+0.26+0.18+0.17+0.02+0.01) = 0.06
b. 0.19+0.26+0.18 = 0.63
c. Check student’s solution.

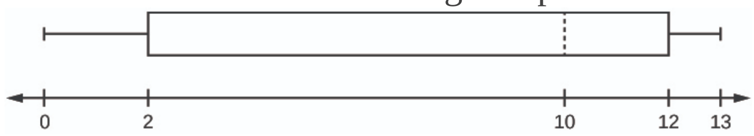
d. 40^{th} percentile will fall between 30,000 and 40,000

   80^{th} percentile will fall between 50,000 and 75,000

e. Check student’s solution.

Exercise:

Problem: Given the following box plot:



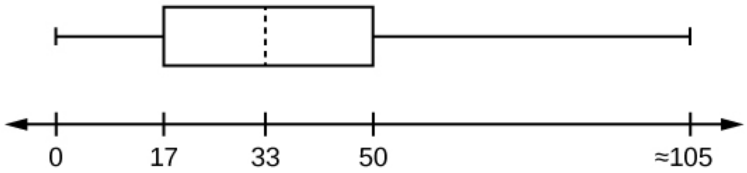
a. which quarter has the smallest spread of data? What is that spread?
b. which quarter has the largest spread of data? What is that spread?
c. find the interquartile range (*IQR*).
d. are there more data in the interval 5–10 or in the interval 10–13? How do you know this?
e. which interval has the fewest data in it? How do you know this?

    i. 0–2
    ii. 2–4
    iii. 10–12
    iv. 12–13
    v. need more information

Exercise:

Problem:

The following box plot shows the U.S. population for 1990, the latest available year.



a. Are there fewer or more children (age 17 and under) than senior citizens (age 65 and over)? How do you know?
b. 12.6% are age 65 and over. Approximately what percentage of the population are working age adults (above age 17 to age 65)?

Solution:

a. more children; the left whisker shows that 25% of the population are children 17 and younger. The right whisker shows that 25% of the population are adults 50 and older, so adults 65 and over represent less than 25%.
b. 62.4%

Glossary

Interquartile Range
: or *IQR*, is the range of the middle 50 percent of the data values; the *IQR* is found by subtracting the first quartile from the third quartile.

Outlier
: an observation that does not fit the rest of the data

Percentile
: a number that divides ordered data into hundredths; percentiles may or may not be part of the data. The median of the data is the second quartile and the 50^{th} percentile. The first and third quartiles are the 25^{th} and the 75^{th} percentiles, respectively.

Quartiles
: the numbers that separate the data into quarters; quartiles may or may not be part of the data. The second quartile is the median of the data.

Box Plots

Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value. We use these values to compare how close other data values are to them.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

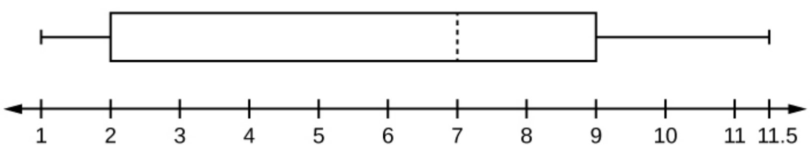
> **Note:**
> NOTE
> You may encounter box-and-whisker plots that have dots marking outlier values. In those cases, the whiskers are not extending to the minimum and maximum values.

Consider, again, this dataset.

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

The first quartile is two, the median is seven, and the third quartile is nine. The smallest value is one, and the largest value is 11.5. The following image shows the constructed box plot.

Note:
NOTE
See the calculator instructions on the [TI web site] or in the appendix.

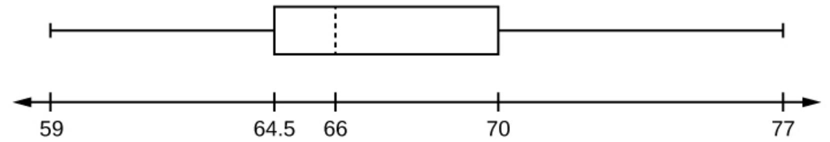


The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

Note:
NOTE
It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

Example:
The following data are the heights of 40 students in a statistics class.
59 60 61 62 62 63 63 64 64 64 65 65 65 65 65 65 65 65 65 66 66 67 67 68 68 69 70 70 70 70 70 71 71 72 72 73 74 74 75 77
Construct a box plot with the following properties; the calculator intructions for the minimum and maximum values as well as the quartiles follow the example.

- Minimum value = 59
- Maximum value = 77
- $Q1$: First quartile = 64.5
- $Q2$: Second quartile or median= 66
- $Q3$: Third quartile = 70



a. Each quarter has approximately 25% of the data.
b. The spreads of the four quarters are 64.5 – 59 = 5.5 (first quarter), 66 – 64.5 = 1.5 (second quarter), 70 – 66 = 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
c. Range = maximum value – the minimum value = 77 – 59 = 18
d. Interquartile Range: $IQR = Q3 - Q1 = 70 - 64.5 = 5.5$.
e. The interval 59–65 has more than 25% of the data so it has more data in it than the interval 66 through 70 which has 25% of the data.
f. The middle 50% (middle half) of the data has a range of 5.5 inches.

Note:
To find the minimum, maximum, and quartiles:
Enter data into the list editor (Pres STAT 1:EDIT). If you need to clear the list, arrow up to the name L1, press CLEAR, and then arrow down.
Put the data values into the list L1.
Press STAT and arrow to CALC. Press 1:1-VarStats. Enter L1.
Press ENTER.
Use the down and up arrow keys to scroll.
Smallest value = 59.
Largest value = 77.
Q_1: First quartile = 64.5.
Q_2: Second quartile or median = 66.
Q_3: Third quartile = 70.

To construct the box plot:
Press 4:Plotsoff. Press ENTER.
Arrow down and then use the right arrow key to go to the fifth picture, which is the box plot. Press ENTER.
Arrow down to Xlist: Press 2nd 1 for L1

Arrow down to Freq: Press ALPHA. Press 1.
Press Zoom. Press 9: ZoomStat.
Press TRACE, and use the arrow keys to examine the box plot.

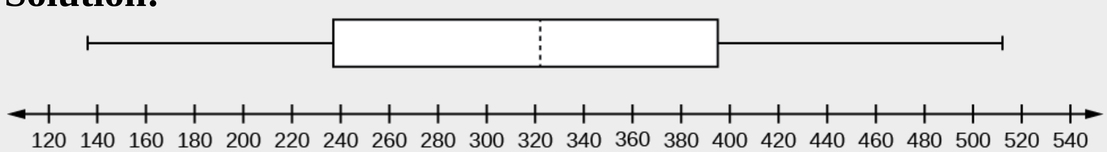
Note:
Try It
Exercise:

Problem:

The following data are the number of pages in 40 books on a shelf. Construct a box plot using a graphing calculator, and state the interquartile range.

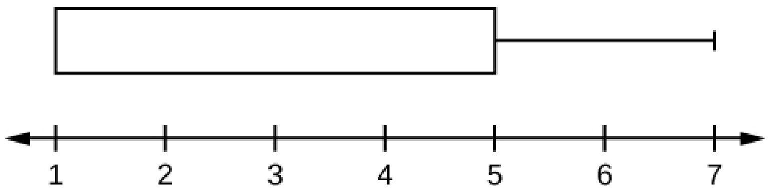
136 140 178 190 205 215 217 218 232 234 240 255 270 275 290 301 303 315 317 318 326 333 343 349 360 369 377 388 391 392 398 400 402 405 408 422 429 450 475 512

Solution:



$IQR = 158$

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Example:
Test scores for a college statistics class held during the day are:
99 56 78 55.5 32 90 80 81 56 59 45 77 84.5 84 70 72 68 32 79 90
Test scores for a college statistics class held during the evening are:
98 78 68 83 81 89 88 76 65 45 98 90 80 84.5 85 79 78 98 90 79 81 25.5
Exercise:

Problem:

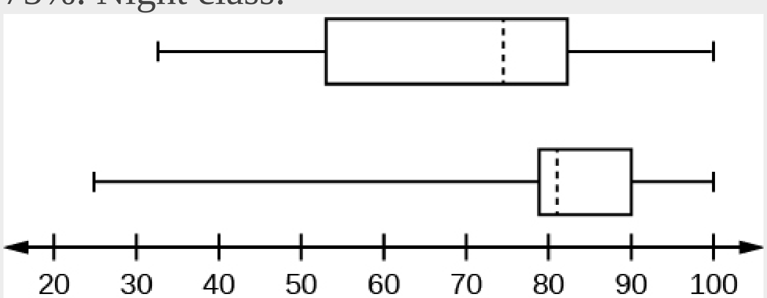
a. Find the smallest and largest values, the median, and the first and third quartile for the day class.
b. Find the smallest and largest values, the median, and the first and third quartile for the night class.
c. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
d. Create a box plot for each set of data. Use one number line for both box plots.
e. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

Solution:

a.
  - Min = 32
  - Q_1 = 56
  - M = 74.5
  - Q_3 = 82.5
  - Max = 99

b.
  - Min = 25.5
  - Q_1 = 78
  - M = 81
  - Q_3 = 89
  - Max = 98

c. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:

d. 

e. The first data set has the wider spread for the middle 50% of the data. The *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

Note:
Try It

Exercise:

Problem:

The following data set shows the heights in inches for the boys in a class of 40 students.

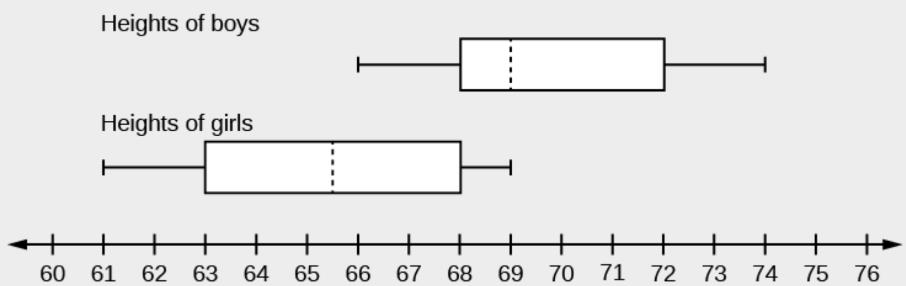
66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using a graphing calculator for each data set, and state which box plot has the wider spread for the middle 50% of the data.

Solution:



IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

Example:

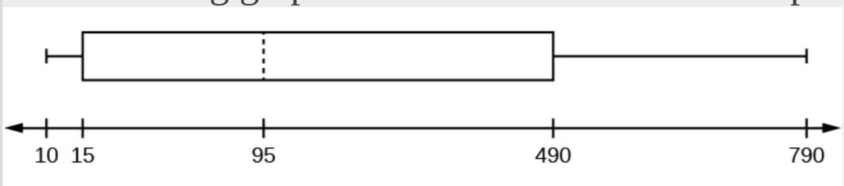
Graph a box-and-whisker plot for the data values shown.

10101015357590951001754204905155157 90

The five numbers used to create a box-and-whisker plot are:

- Min: 10
- Q_1: 15
- Med: 95
- Q_3: 490
- Max: 790

The following graph shows the box-and-whisker plot.



Note:
Try It
Exercise:

Problem:

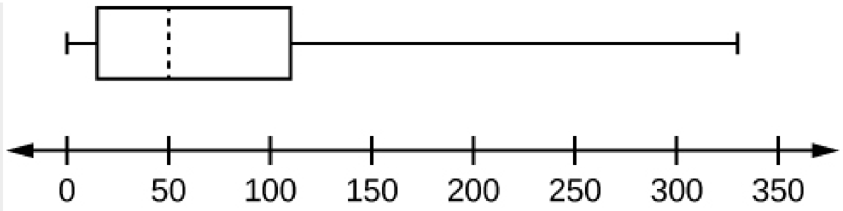
Follow the steps you used to graph a box-and-whisker plot for the data values shown.

055153030455050607511014024033O

Solution:

The data are in order from least to greatest. There are 15 values, so the eighth number in order is the median: 50. There are seven data values written to the left of the median and 7 values to the right. The five values that are used to create the boxplot are:

- Min: 0
- Q_1: 15
- Med: 50
- Q_3: 110
- Max: 330



References

Data from *West Magazine*.

Chapter Review

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

Use the following information to answer the next two exercises. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars.

Exercise:

Problem:

Construct a box plot below. Use a ruler to measure and scale accurately.

Exercise:

Problem:

Looking at your box plot, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

Solution:

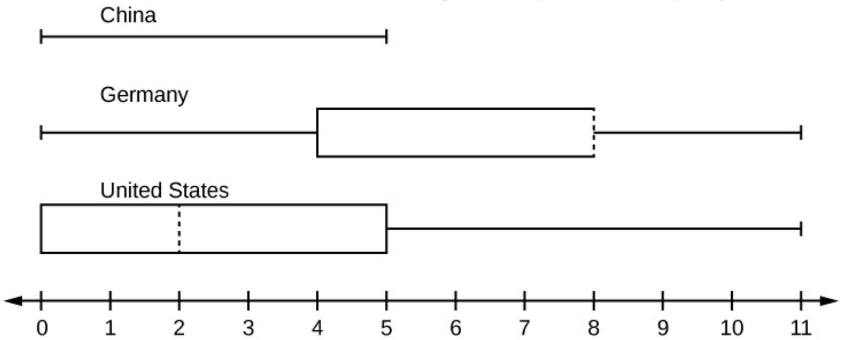
More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

Homework

Exercise:

Problem:

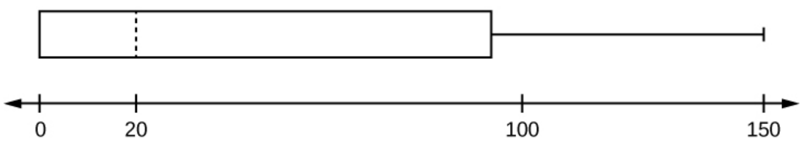
In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
b. Have more Americans or more Germans surveyed been to over eight foreign countries?
c. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

Exercise:

Problem: Given the following box plot, answer the questions.



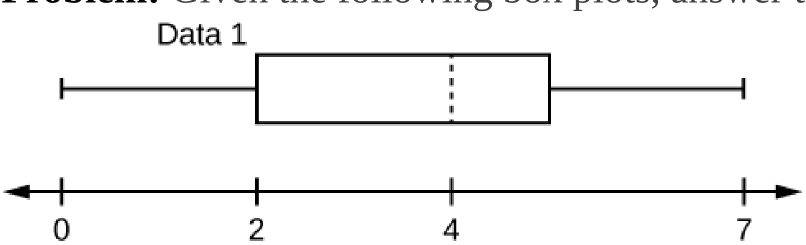
a. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.
b. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

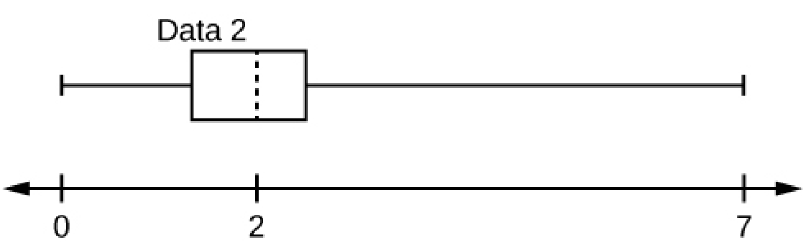
Solution:

a. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.
b. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

Exercise:

Problem: Given the following box plots, answer the questions.





a. In complete sentences, explain why each statement is false.

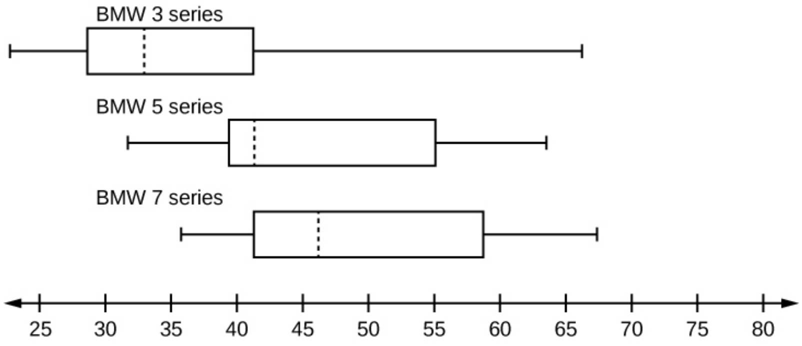
   i. **Data 1** has more data values above two than **Data 2** has above two.
   ii. The data sets cannot have the same mode.
   iii. For **Data 1**, there are more data values below four than there are above four.

b. For which group, Data 1 or Data 2, is the value of "7" more likely to be an outlier? Explain why in complete sentences.

Exercise:

Problem:

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



a. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
b. Which group is most likely to have an outlier? Explain how you determined that.
c. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
d. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?

e. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
f. Look at the BMW 5 series. Estimate the interquartile range (IQR).
g. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
h. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

    i. 31–35
    ii. 38–41
    iii. 41–64

Solution:

a. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.
b. The BMW 3 series is most likely to have an outlier. It has the longest whisker.
c. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
d. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
e. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
f. $IQR \sim 17$ years
g. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.
h. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and

25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4
4	1

Construct a box plot of the data.

Bringing It Together

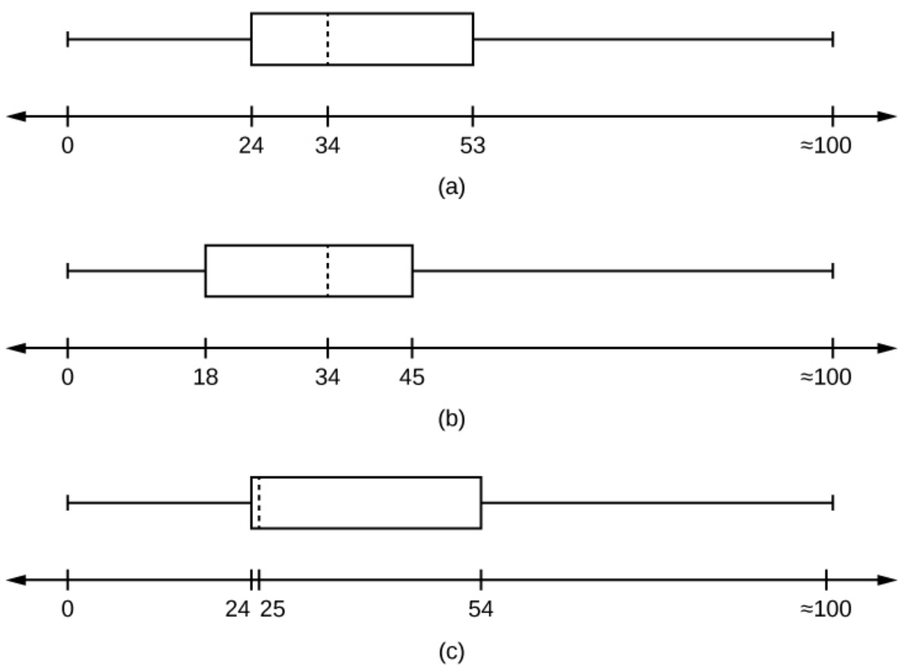
Exercise:

Problem:

Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

a. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
b. What percentage of the community is under age 35?
c. Which box plot most resembles the information above?



Solution:

a. For graph, check student's solution.
b. 49.7% of the community is under the age of 35.
c. Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot
: a graph that gives a quick picture of the middle 50% of the data

First Quartile
: the value that is the median of the of the lower half of the ordered data set

Frequency Polygon
: looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

Measures of the Center of the Data

The "center" of a data set is also a way of describing location. The two most widely used measures of the "center" of the data are the **mean** (average) and the **median**. To calculate the **mean weight** of 50 people, add the 50 weights together and divide by 50. To find the **median weight** of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

Note:
NOTE
The words “mean” and “average” are often used interchangeably. The substitution of one word for the other is common practice. The technical term is “arithmetic mean” and “average” is technically a center location. However, in practice among non-statisticians, “average" is commonly accepted for “arithmetic mean.”

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. The letter used to represent the **sample mean** is an *x* with a bar over it (pronounced “*x* bar”): x.

The Greek letter μ (pronounced "mew") represents the **population mean**. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

To see that both ways of calculating the mean are the same, consider the sample:
1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4
Equation:

$$x = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.7$$

Equation:

$$x = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.7$$

In the second calculation, the frequencies are 3, 2, 1, and 5.

You can quickly find the location of the median by using the expression $\frac{n+1}{2}$.

The letter n is the total number of data values in the sample. If n is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then $\frac{n+1}{2} = \frac{97+1}{2} = 49$. The median is the 49th value in the ordered data. If the total number of data values is 100, then $\frac{n+1}{2} = \frac{100+1}{2} = 50.5$. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter M is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Example:
Exercise:

Problem:

AIDS data indicating the number of months a patient with AIDS lives after taking a new antibody drug are as follows (smallest to largest):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;
Calculate the mean and the median.

Solution:

The calculation for the mean is:

$$x = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+...+35+37+40+(44)(2)+47]}{40} = 23.6$$

To find the median, *M*, first use the formula for the location. The location is:
$\frac{n+1}{2} = \frac{40+1}{2} = 20.5$
Starting at the smallest value, the median is located between the 20^{th} and 21^{st} values (the two 24s):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24+24}{2} = 24$$

Note:
To find the mean and the median:
Clear list L1. Pres STAT 4:ClrList. Enter 2nd 1 for list L1. Press ENTER.
Enter data into the list editor. Press STAT 1:EDIT.
Put the data values into list L1.
Press STAT and arrow to CALC. Press 1:1-VarStats. Press 2nd 1 for L1 and then ENTER.
Press the down and up arrow keys to scroll.
x = 23.6, M = 24

Note:
Try It
Exercise:

Problem:

The following data show the number of months patients typically wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3 4 5 7 7 7 7 8 8 9 9 10 10 10 10 10 11 12 12 13 14 14 15 15 17 17 18 19 19 19 21 21 22 22 23 24 24 24 24

Solution:

Mean: 3 + 4 + 5 + 7 + 7 + 7 + 7 + 8 + 8 + 9 + 9 + 10 + 10 + 10 + 10 + 10 + 11 + 12 + 12 + 13 + 14 + 14 + 15 + 15 + 17 + 17 + 18 + 19 + 19 + 19 + 21 + 21 + 22 + 22 + 23 + 24 + 24 + 24 = 544
$\frac{544}{39} = 13.95$
Median: Starting at the smallest value, the median is the 20th term, which is 13.

Example:
Exercise:

Problem:

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean or the median?

Solution:

$x = \frac{5{,}000{,}000+49(30{,}000)}{50} = 129{,}400$

M = 30,000

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The median is a better measure of the "center" than the mean because 49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

Note:
Try It
Exercise:

Problem:

In a sample of 60 households, one house is worth $2,500,000. Half of the rest are worth $280,000, and all the others are worth $315,000. Which is the better measure of the “center”: the mean or the median?

Solution:

The median is the better measure of the “center” than the mean because 59 of the values are $280,000 and one is $2,500,000. The $2,500,000 is an outlier. Either $280,000 or $315,000 gives us a better sense of the middle of the data.

Another measure of the center is the mode. The **mode** is the most frequent value. There can be more than one mode in a data set as long as those values have the same frequency and that frequency is the highest. A data set with two modes is called bimodal.

Example:
Statistics exam scores for 20 students are as follows:
50535959636372727272727678818384848493
Exercise:

Problem: Find the mode.

Solution:

The most frequent score is 72, which occurs five times. Mode = 72.

Note:
Try It
Exercise:

Problem: The number of books checked out from the library from 25 students are as follows:

0001233445577778889101011111212
Find the mode.

Solution:

The most frequent number of books is 7, which occurs four times. Mode = 7.

Example:
Five real estate exam scores are 430, 430, 480, 480, 495. The data set is bimodal because the scores 430 and 480 each occur twice.
When is the mode the best measure of the "center"? Consider a weight loss program that advertises a mean weight loss of six pounds the first week of the program. The mode might indicate that most people lose two pounds the first week, making the program less appealing.

Note:
NOTE
The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

Statistical software will easily calculate the mean, the median, and the mode. Some graphing calculators can also make these calculations. In the real world, people make these calculations using software.

Note:
Try It

Exercise:

Problem:

Five credit scores are 680, 680, 700, 720, 720. The data set is bimodal because the scores 680 and 720 each occur twice. Consider the annual earnings of workers at a factory. The mode is $25,000 and occurs 150 times out of 301. The median is $50,000 and the mean is $47,500. What would be the best measure of the “center”?

Solution:

Because $25,000 occurs nearly half the time, the mode would be the best measure of the center because the median and mean don’t represent what most people make at the factory.

The Law of Large Numbers and the Mean

The Law of Large Numbers says that if you take samples of larger and larger size from any population, then the mean x of the sample is very likely to get closer and closer to μ. This is discussed in more detail later in the text.

Sampling Distributions and Statistic of a Sampling Distribution

You can think of a **sampling distribution** as a **relative frequency distribution** with a great many samples. (See **Sampling and Data** for a review of relative frequency). Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are in the **relative frequency table** shown below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$

# of movies	Relative Frequency
3	$\frac{3}{30}$
4	$\frac{1}{30}$

If you let the number of samples get very large (say, 300 million or more), the relative frequency table becomes a relative frequency distribution.

A **statistic** is a number calculated from a sample. Statistic examples include the mean, the median and the mode as well as others. The sample mean x is an example of a statistic which estimates the population mean μ.

Calculating the Mean of Grouped Frequency Tables

When only grouped data is available, you do not know the individual data values (we only know intervals and interval frequencies); therefore, you cannot compute an exact mean for the data set. What we must do is estimate the actual mean by calculating the mean of a frequency table. A frequency table is a data representation in which grouped data is displayed along with the corresponding frequencies. To calculate the mean from a grouped frequency table we can apply the basic definition of mean: *mean* = $\frac{data\ sum}{number\ of\ data\ values}$ We simply need to modify the definition to fit within the restrictions of a frequency table.

Since we do not know the individual data values we can instead find the midpoint of each interval. The midpoint is $\frac{lower\ boundary + upper\ boundary}{2}$. We can now modify the mean definition to be $Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$ where *f* = the frequency of the interval and *m* = the midpoint of the interval.

Example:
Exercise:

Problem:

A frequency table displaying professor Blount’s last statistic test is shown. Find the best estimate of the class mean.

Grade Interval	Number of Students

Grade Interval	Number of Students
50–56.5	1
56.5–62.5	0
62.5–68.5	4
68.5–74.5	4
74.5–80.5	2
80.5–86.5	3
86.5–92.5	4
92.5–98.5	1

Solution:

- Find the midpoints for all intervals

Grade Interval	Midpoint
50–56.5	53.25
56.5–62.5	59.5
62.5–68.5	65.5
68.5–74.5	71.5
74.5–80.5	77.5
80.5–86.5	83.5
86.5–92.5	89.5
92.5–98.5	95.5

- Calculate the sum of the product of each interval frequency and midpoint. $\sum fm$

  $53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$

- $\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$

Note:
Try It
Exercise:

Problem:

Maris conducted a study on the effect that playing video games has on memory recall. As part of her study, she compiled the following data:

Hours Teenagers Spend on Video Games	**Number of Teenagers**
0–3.5	3
3.5–7.5	7
7.5–11.5	12
11.5–15.5	7
15.5–19.5	9

What is the best estimate for the mean number of hours spent playing video games?

Solution:

Find the midpoint of each interval, multiply by the corresponding number of teenagers, add the results and then divide by the total number of teenagers
The midpoints are 1.75, 5.5, 9.5, 13.5,17.5.
Mean $= \frac{(1.75)(3) + (5.5)(7) + (9.5)(12) + (13.5)(7) + (17.5)(9)}{(3 + 7 + 12 + 7 + 9)} = \frac{409.75}{38} = 10.78$

References

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Chapter Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

Formula Review

$\mu = \frac{\sum fm}{\sum f}$ Where *f* = interval frequencies and *m* = interval midpoints.

Exercise:

Problem: Find the mean for the following frequency tables.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1

Daily Low Temperature	Frequency
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Use the following information to answer the next three exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest:
161719202021232425252526262727272829303233333435373940

Exercise:

Problem: Calculate the mean.

Solution:

Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738;

$\frac{738}{27}$ = 27.33

Exercise:

Problem: Identify the median.

Exercise:

Problem: Identify the mode.

Solution:

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three exercises: Sixty-five randomly selected car

salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

Exercise:

Problem: sample mean = $\bar{x}$ = ________

Exercise:

Problem: median = ________

Solution:

4

Exercise:

Problem: mode = ________

Homework

Exercise:

Problem:

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in the following table.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

a. What is the best estimate of the average obesity percentage for these countries?
b. The United States has an average obesity rate of 33.9%. Is this rate above average or below?
c. How does the United States compare to other countries?

Exercise:

Problem:

[link] gives the percent of children under five considered to be underweight. What is the best estimate for the mean percentage of underweight children?

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

Solution:

The mean percentage, $x = \frac{1328.65}{50} = 26.75$

Bringing It Together

Exercise:

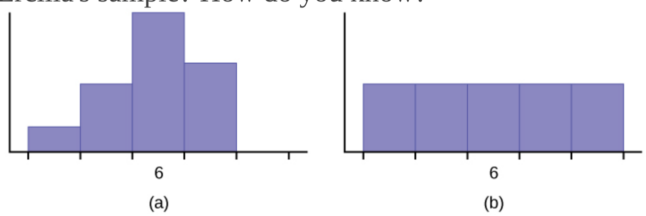
Problem:

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

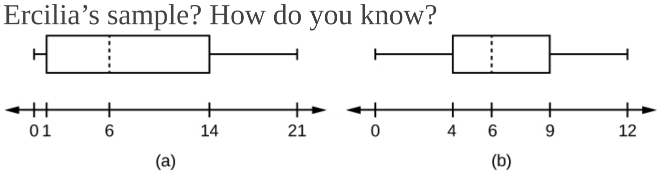
	Javier	Ercilia

	Javier	Ercilia
x	6.0 miles	6.0 miles
s	4.0 miles	7.0 miles

a. How can you determine which survey was correct ?
b. Explain what the difference in the results of the surveys implies about the data.
c. If the two histograms depict the distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



d. If the two box plots depict the distribution of values for each supervisor, which one depicts Ercilia’s sample? How do you know?



Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	Frequency	Number of years	Frequency
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		
			Total = 20

Exercise:

Problem: What is the *IQR*?

a. 8
b. 11
c. 15
d. 35

Solution:

a

Exercise:

Problem: What is the mode?

a. 19
b. 19.5
c. 14 and 20
d. 22.65

Exercise:

Problem: Is this a sample or the entire population?

a. sample
b. entire population
c. neither

Solution:

b

Glossary

Frequency Table
: a data representation in which grouped data is displayed along with the corresponding frequencies

Mean
: a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x) is $x\ =\ \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Median
: a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may

or may not be part of the data.

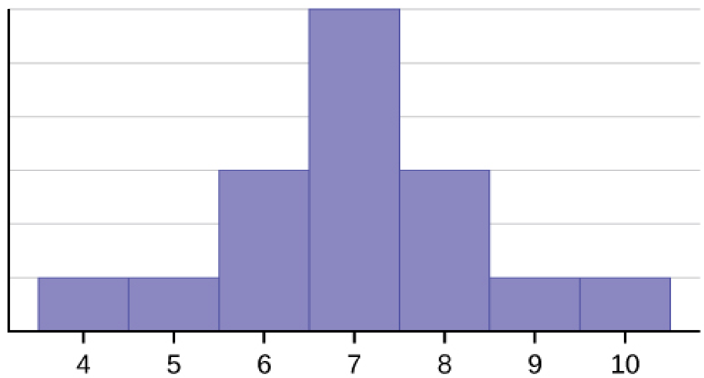
Midpoint
the mean of an interval in a frequency table

Mode
the value that appears most frequently in a set of data

Skewness and the Mean, Median, and Mode

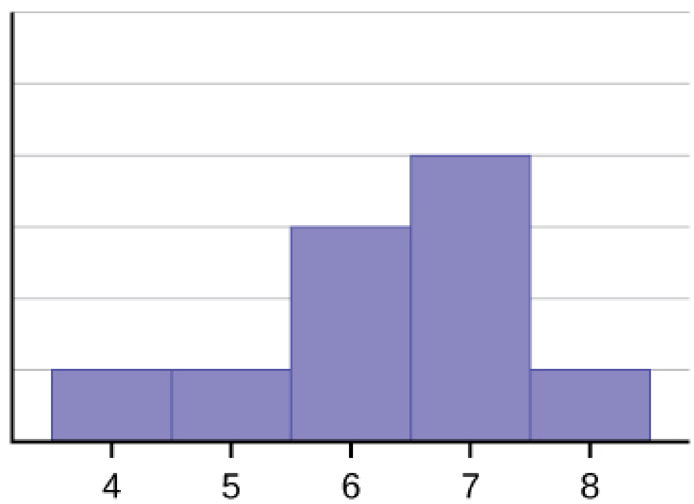
Consider the following data set.
4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

This data set can be represented by following histogram. Each interval has width one, and each value is located in the middle of an interval.



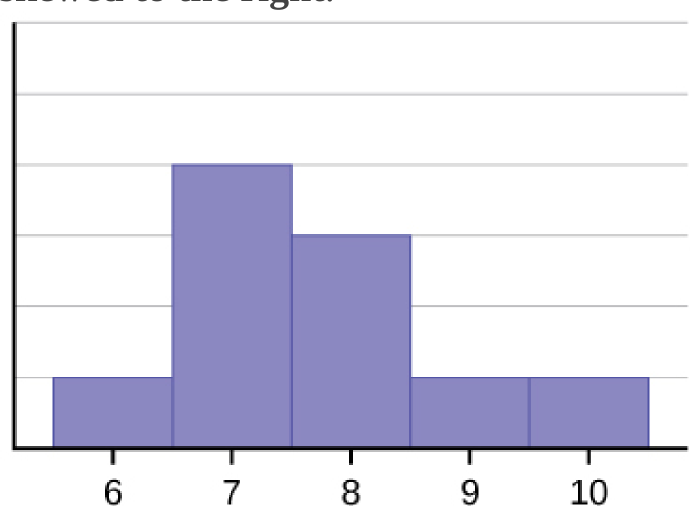
The histogram displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the histogram such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data. **In a perfectly symmetrical distribution, the mean and the median are the same.** This example has one mode (unimodal), and the mode is the same as the mean and median. In a symmetrical distribution that has two modes (bimodal), the two modes would be different from the mean and median.

The histogram for the data: 4566677778 is not symmetrical. The right-hand side seems "chopped off" compared to the left side. A distribution of this type is called **skewed to the left** because it is pulled out to the left.



The mean is 6.3, the median is 6.5, and the mode is seven. **Notice that the mean is less than the median, and they are both less than the mode.** The mean and the median both reflect the skewing, but the mean reflects it more so.

The histogram for the data: 67777888910, is also not symmetrical. It is **skewed to the right**.



The mean is 7.7, the median is 7.5, and the mode is seven. Of the three statistics, **the mean is the largest, while the mode is the smallest**. Again, the mean reflects the skewing the most.

To summarize, generally if the distribution of data is skewed to the left, the mean is less than the median, which is often less than the mode. If the

distribution of data is skewed to the right, the mode is often less than the median, which is less than the mean.

Skewness and symmetry become important when we discuss probability distributions in later chapters.

Example:
Exercise:

Problem:

Statistics are used to compare and sometimes identify authors. The following lists shows a simple random sample that compares the letter counts for three authors.

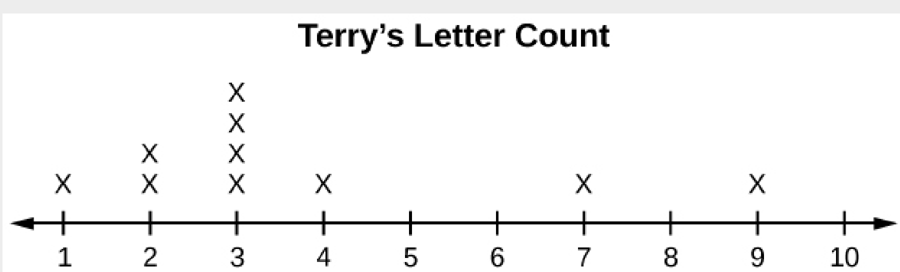
Terry: 7; 9; 3; 3; 3; 4; 1; 3; 2; 2

Davis: 3; 3; 3; 4; 1; 4; 3; 2; 3; 1

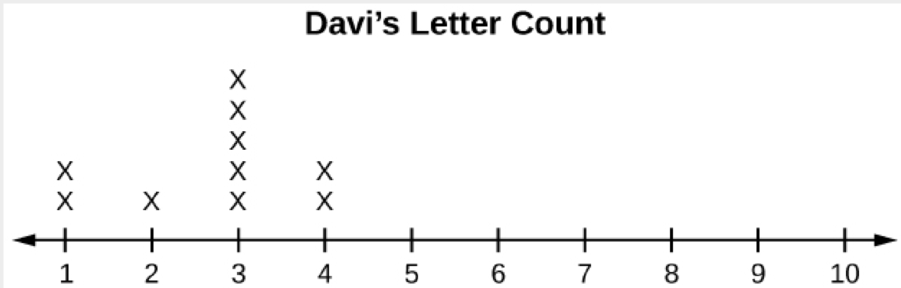
Maris: 2; 3; 4; 4; 4; 6; 6; 6; 8; 3

a. Make a dot plot for the three authors and compare the shapes.
b. Calculate the mean for each.
c. Calculate the median for each.
d. Describe any pattern you notice between the shape and the measures of center.

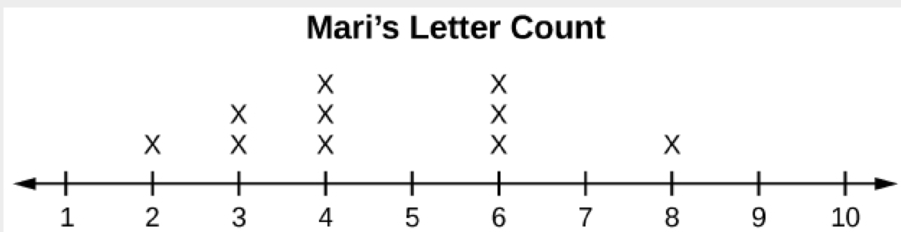
Solution:

a. 

Terry’s distribution has a right (positive) skew.



Davis’ distribution has a left (negative) skew



Maris’ distribution is symmetrically shaped.

b. Terry’s mean is 3.7, Davis’ mean is 2.7, Maris’ mean is 4.6.
c. Terry’s median is three, Davis’ median is three. Maris’ median is four.
d. It appears that the median is always closest to the high point (the mode), while the mean tends to be farther out on the tail. In a symmetrical distribution, the mean and the median are both centrally located close to the high point of the distribution.

Note:
Try It
Exercise:

Problem:

Discuss the mean, median, and mode for each of the following problems. Is there a pattern between the shape and measure of the center?

a.

b.

The Ages Former U.S Presidents Died	
4	6 9
5	3 6 7 7 7 8
6	0 0 3 3 4 4 5 6 7 7 7 8
7	0 1 1 2 3 4 7 8 8 9
8	0 1 3 5 8
9	0 0 3 3
Key: 8|0 means 80.	

c.

Solution:

a. mean = 4.25, median = 3.5, mode = 1; The mean > median > mode which indicates skewness to the right. (data are 0, 1, 2, 3, 4, 5, 6, 9, 10, 14 and respective frequencies are 2, 4, 3, 1, 2, 2, 2, 2, 1, 1)
b. mean = 70.1 , median = 68, mode = 57, 67 bimodal; the mean and median are close but there is a little skewness to the right which is influenced by the data being bimodal. (data are 46, 49, 53, 56, 57, 57, 57, 58, 60, 60, 63, 63, 64, 64, 65, 66, 67, 67, 67, 68, 70, 71, 71, 72, 73, 74, 77, 78, 78, 79, 80, 81, 83, 85, 88, 90, 90 93, 93).
c. These are estimates: mean =16.095, median = 17.495, mode = 22.495 (there may be no mode); The mean < median < mode which indicates skewness to the left. (data are the midponts of the intervals: 2.495, 7.495, 12.495, 17.495, 22.495 and respective frequencies are 2, 3, 4, 7, 9).

Chapter Review

Looking at the distribution of data can reveal a lot about the relationship between the mean, the median, and the mode. There are three types of

distributions. A **right (or positive) skewed** distribution has a shape like [link]. A **left (or negative) skewed** distribution has a shape like [link]. A **symmetrical** distrubtion looks like [link].

Use the following information to answer the next three exercises: State whether the data are symmetrical, skewed to the left, or skewed to the right.

Exercise:

Problem: 1 1 1 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 5 5

Solution:

The data are symmetrical. The median is 3 and the mean is 2.85. They are close, and the mode lies close to the middle of the data, so the data are symmetrical.

Exercise:

Problem:16 17 19 22 22 22 22 22 23

Exercise:

Problem:87 87 87 87 87 88 89 89 90 91

Solution:

The data are skewed right. The median is 87.5 and the mean is 88.2. Even though they are close, the mode lies to the left of the middle of the data, and there are many more instances of 87 than any other number, so the data are skewed right.

Exercise:

Problem:

When the data are skewed left, what is the typical relationship between the mean and median?

Exercise:

Problem:

When the data are symmetrical, what is the typical relationship between the mean and median?

Solution:

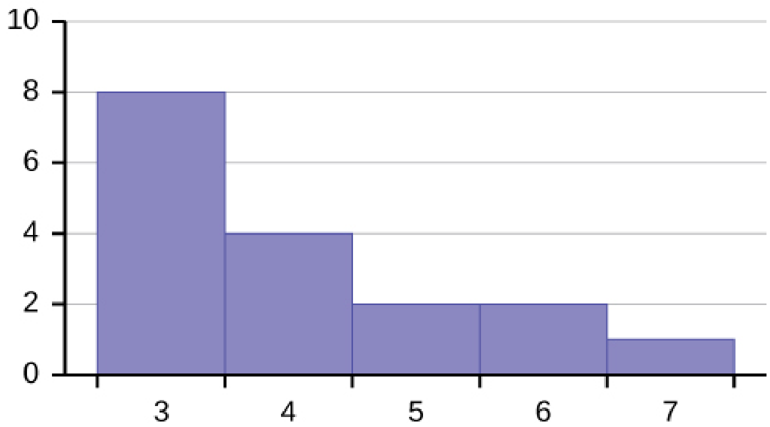
When the data are symmetrical, the mean and median are close or the same.

Exercise:

Problem: What word describes a distribution that has two modes?

Exercise:

Problem: Describe the shape of this distribution.



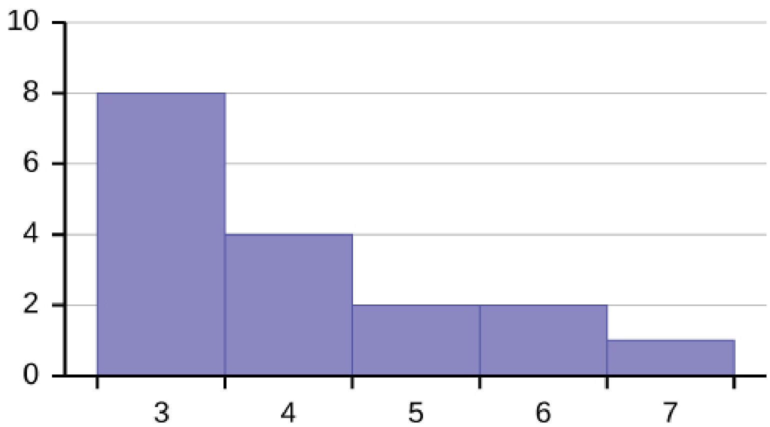
Solution:

The distribution is skewed right because it looks pulled out to the right.

Exercise:

Problem:

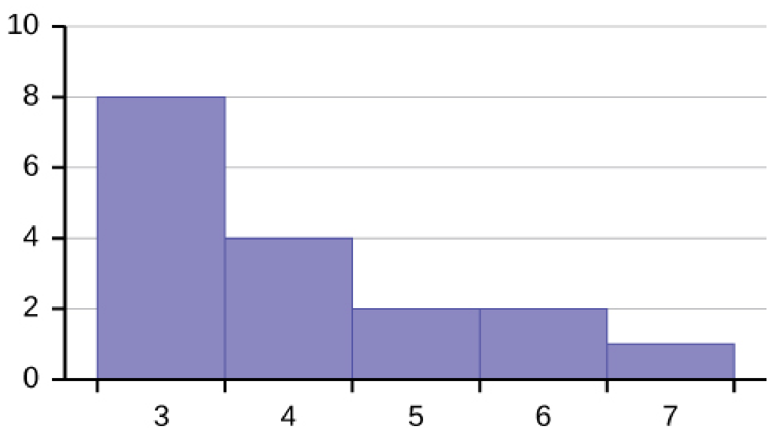
Describe the relationship between the mode and the median of this distribution.



Exercise:

Problem:

Describe the relationship between the mean and the median of this distribution.

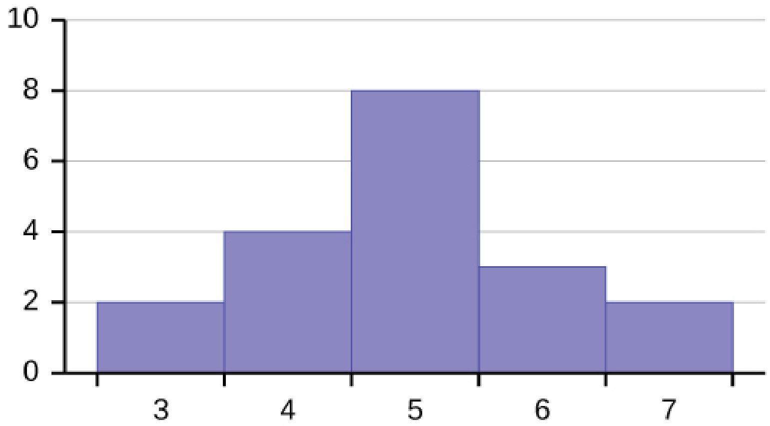


Solution:

The mean is 4.1 and is slightly greater than the median, which is four.

Exercise:

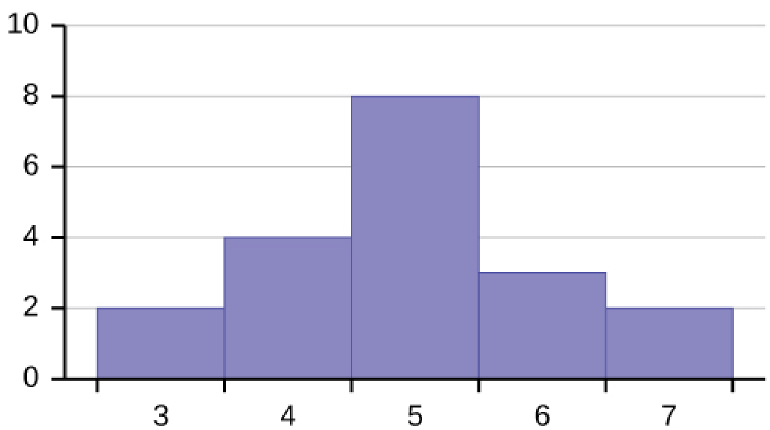
Problem: Describe the shape of this distribution.



Exercise:

Problem:

Describe the relationship between the mode and the median of this distribution.



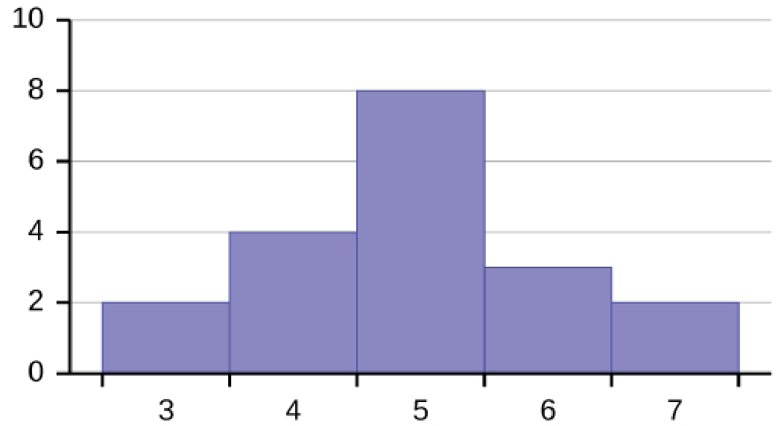
Solution:

The mode and the median are the same. In this case, they are both five.

Exercise:

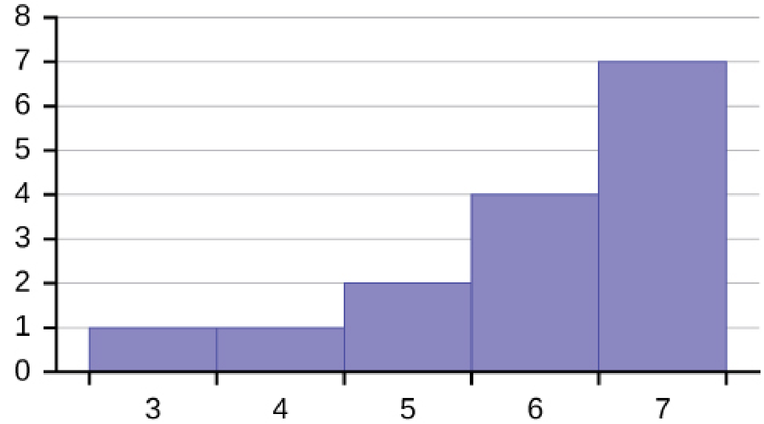
Problem:

Are the mean and the median the exact same in this distribution? Why or why not?



Exercise:

Problem: Describe the shape of this distribution.



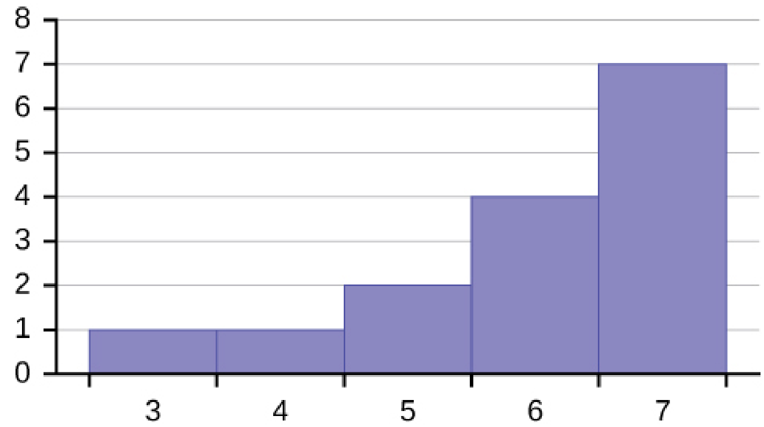
Solution:

The distribution is skewed left because it looks pulled out to the left.

Exercise:

Problem:

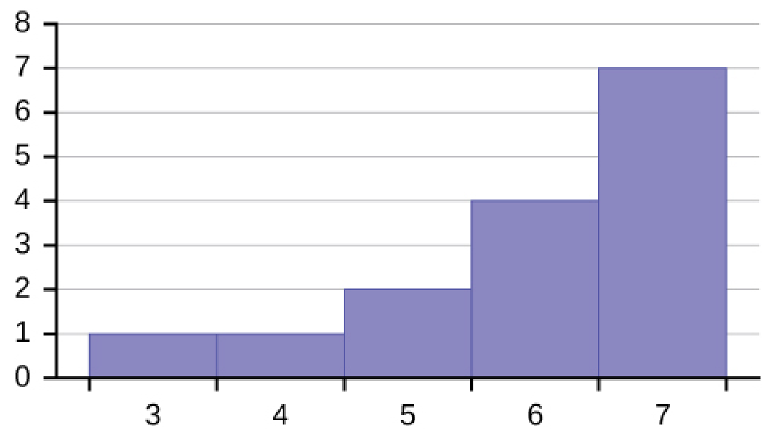
Describe the relationship between the mode and the median of this distribution.



Exercise:

Problem:

Describe the relationship between the mean and the median of this distribution.



Solution:

The mean and the median are both six.

Exercise:

Problem: The mean and median for the data are the same.

345566667777777

Is the data perfectly symmetrical? Why or why not?

Exercise:

Problem:

Which is the greatest, the mean, the mode, or the median of the data set?

11111212121212131517222222

Solution:

The mode is 12, the median is 12.5, and the mean is 15.1. The mean is the largest.

Exercise:

Problem:

Which is the least, the mean, the mode, and the median of the data set?

56565658596062646567

Exercise:

Problem:

Of the three measures, which tends to reflect skewing the most, the mean, the mode, or the median? Why?

Solution:

The mean tends to reflect skewing the most because it is affected the most by outliers.

Exercise:

Problem:

In a perfectly symmetrical distribution, when would the mode be different from the mean and median?

Homework

Exercise:

Problem:

The median age of the U.S. population in 1980 was 30.0 years. In 1991, the median age was 33.1 years.

a. What does it mean for the median age to rise?
b. Give two reasons why the median age could rise.
c. For the median age to rise, is the actual number of children less in 1991 than it was in 1980? Why or why not?

Measures of the Spread of the Data

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation, or spread, is the standard deviation. The **standard deviation** is a number that measures how far data values are from their mean.

The standard deviation

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket *A* and supermarket *B*. the average wait time at both supermarkets is five minutes. At supermarket *A*, the standard deviation for the wait time is two minutes; at supermarket *B* the standard deviation for the wait time is four minutes.

Because supermarket *B* has a higher standard deviation, we know that there is more variation in the wait times at supermarket *B*. Overall, wait times at supermarket *B* are more spread out from the average; wait times at supermarket *A* are more concentrated near the average.

The standard deviation can be used to determine whether a data value is close to or far from the mean.

Suppose that Rosa and Binh both shop at supermarket *A*. Rosa waits at the checkout counter for seven minutes and Binh waits for one minute. At supermarket *A*, the mean waiting time is five minutes and the standard deviation is two minutes. The standard deviation can be used to determine whether a data value is close to or far from the mean.

Rosa waits for seven minutes:

- Seven is two minutes longer than the average of five; two minutes is equal to one standard deviation.
- Rosa's wait time of seven minutes is **two minutes longer than the average** of five minutes.
- Rosa's wait time of seven minutes is **one standard deviation above the average** of five minutes.

Binh waits for one minute.

- One is four minutes less than the average of five; four minutes is equal to two standard deviations.
- Binh's wait time of one minute is **four minutes less than the average** of five minutes.
- Binh's wait time of one minute is **two standard deviations below the average** of five minutes.
- A data value that is two standard deviations from the average is just on the borderline for what many statisticians would consider to be far from the average. Considering data to be far from the mean if it is more than two standard deviations away is more of an approximate "rule of thumb" than a rigid rule. In general, the shape of the distribution of the data affects how much of the data is further away than two standard deviations. (You will learn more about this in later chapters.)

The number line may help you understand standard deviation. If we were to put five and seven on a number line, seven is to the right of five. We say, then, that seven is **one** standard deviation to the **right** of five because 5 + (1)(2) = 7.

If one were also part of the data set, then one is **two** standard deviations to the **left** of five because 5 + (–2)(2) = 1.

- In general, a **value = mean + (#ofSTDEV)(standard deviation)**
- where #ofSTDEVs = the number of standard deviations
- #ofSTDEV does not need to be an integer
- One is **two standard deviations less than the mean** of five because: 1 = 5 + (–2)(2).

The equation **value = mean + (#ofSTDEVs)(standard deviation)** can be expressed for a sample and for a population.

- **sample:** $x = x + (\#ofSTDEV)(s)$
- **Population:** $x = \mu + (\#ofSTDEV)(\sigma)$

The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation.

The symbol x is the sample mean and the Greek symbol μ is the population mean.

Calculating the Standard Deviation

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is $x - \mu$. For sample data, in symbols a deviation is $x - x$.

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the $x - x$ values for a sample, or the $x - \mu$ values for a population). The symbol σ^2 represents the population variance; the population standard deviation σ is the square root of the population variance. The symbol s^2 represents the sample variance; the sample standard deviation s is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by N, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by ***n* – 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s=\sqrt{\frac{\Sigma(x-x)^2}{n-1}}$ or $s=\sqrt{\frac{\Sigma f(x-x)^2}{n-1}}$
- For the sample standard deviation, the denominator is ***n* - 1**, that is the sample size MINUS 1.

Formulas for the Population Standard Deviation

- $\sigma\ =\ \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ or $\sigma\ =\ \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$
- For the population standard deviation, the denominator is *N*, the number of items in the population.

In these formulas, *f* represents the frequency with which a value appears. For example, if a value appears once, *f* is one. If a value appears three times in the data set or population, *f* is three.

Sampling Variability of a Statistic

The statistic of a sampling distribution was discussed in Descriptive Statistics: Measuring the Center of the Data. How much the statistic varies from one sample to another is known as the **sampling variability of a statistic**. You typically measure the sampling variability of a statistic by its standard error. The **standard error of the mean** is an example of a standard error. It is a special standard deviation and is known as the standard deviation of the sampling distribution of the mean. You will cover the standard error of the mean in the chapter The Central Limit Theorem (not now). The notation for the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population and n is the size of the sample.

Note:
NOTE
In practice, USE A CALCULATOR OR COMPUTER SOFTWARE TO CALCULATE THE STANDARD DEVIATION. If you are using a TI-83, 83+, 84+ calculator, you need to select the appropriate standard deviation σ_x or s_x from the summary statistics. We will concentrate on using and interpreting the information that the standard deviation gives us. However you should study the following step-by-step example to help you understand how the standard deviation measures variation from the mean. (The calculator instructions appear at the end of this example.)

Example:
In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of n = 20 fifth grade students. The ages are rounded to the nearest half year:
9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;
Equation:

$$x = \frac{9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)}{20} = 10.525$$

The average age is 10.53 years, rounded to two places.
The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating s.

Data	Freq.	Deviations	*Deviations*^{2}	**(Freq.) (*Deviations*^{2})**
x	f	$(x – x)$	$(x – x)^2$	$(f)(x – x)^2$
9	1	9 – 10.525 = –1.525	$(–1.525)^2$ = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = –0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = –0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875

Data	Freq.	Deviations	*Deviations*^{2}	(Freq.)(*Deviations*^{2})
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number of data values minus one (20 – 1):

$s^2 = \frac{9.7375}{20-1} = 0.5125$

The **sample standard deviation** *s* is equal to the square root of the sample variance:

$s = \sqrt{0.5125} = 0.715891$, which is rounded to two decimal places, *s* = 0.72.

Typically, you do the calculation for the standard deviation on your calculator or computer. The intermediate results are not rounded. This is done for accuracy.

Exercise:

Problem:

- For the following problems, recall that **value = mean + (#ofSTDEVs)(standard deviation)**. Verify the mean and standard deviation or a calculator or computer.
- For a sample: $x = \bar{x}$ + (#ofSTDEVs)(s)
- For a population: $x = \mu$ + (#ofSTDEVs)(σ)
- For this example, use $x = \bar{x}$ + (#ofSTDEVs)(s) because the data is from a sample

a. Verify the mean and standard deviation on your calculator or computer.
b. Find the value that is one standard deviation above the mean. Find ($\bar{x}$ + 1s).
c. Find the value that is two standard deviations below the mean. Find ($\bar{x}$ – 2s).
d. Find the values that are 1.5 standard deviations **from** (below and above) the mean.

Solution:

a. **Note:**

   - Clear lists L1 and L2. Press STAT 4:ClrList. Enter 2nd 1 for L1, the comma (,), and 2nd 2 for L2.
   - Enter data into the list editor. Press STAT 1:EDIT. If necessary, clear the lists by arrowing up into the name. Press CLEAR and arrow down.

- Put the data values (9, 9.5, 10, 10.5, 11, 11.5) into list L1 and the frequencies (1, 2, 4, 4, 6, 3) into list L2. Use the arrow keys to move around.
- Press STAT and arrow to CALC. Press 1:1-VarStats and enter L1 (2nd 1), L2 (2nd 2). Do not forget the comma. Press ENTER.
- x = 10.525
- Use Sx because this is sample data (not a population): Sx=0.715891

b. (x + 1s) = 10.53 + (1)(0.72) = 11.25
c. (x – 2s) = 10.53 – (2)(0.72) = 9.09

d.
- (x – 1.5s) = 10.53 – (1.5)(0.72) = 9.45
- (x + 1.5s) = 10.53 + (1.5)(0.72) = 11.61

Note:
Try It
Exercise:

Problem: On a baseball team, the ages of each of the players are as follows:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Use your calculator or computer to find the mean and standard deviation. Then find the value that is two standard deviations above the mean.

Solution:

μ = 30.68

s = 6.09
(x + 2s) = 30.68 + (2)(6.09) = 42.86.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For [link], there are n = 20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by n = 20, the calculation divided by n – 1 = 20 – 1 = 19 because the data is a sample. For the **sample** variance, we divide by the sample size minus one (n – 1). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by (n – 1) gives a better estimate of the population variance.

Note:
NOTE
Your concentration should be on what the standard deviation tells us about the data. The standard deviation is a number which measures how far the data are spread from the mean. Let a calculator or computer do the arithmetic.

The standard deviation, s or σ, is either zero or larger than zero. Describing the data with reference to the spread is called "variability". The variability in data depends upon the method by which the outcomes are obtained; for example, by measuring or by random sampling. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

The standard deviation, when first presented, can seem unclear. By graphing your data, you can get a better "feel" for the deviations and the standard deviation. You will find that in symmetrical distributions, the standard deviation can be very helpful but in skewed distributions, the standard deviation may not be much help. The reason is that

the two sides of a skewed distribution have different spreads. In a skewed distribution, it is better to look at the first quartile, the median, the third quartile, the smallest value, and the largest value. Because numbers can be confusing, **always graph your data**. Display your data in a histogram or a box plot.

Example:
Exercise:

Problem:

Use the following data (first exam scores) from Susan Dean's spring pre-calculus class:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

a. Create a chart containing the data, frequencies, relative frequencies, and cumulative relative frequencies to three decimal places.
b. Calculate the following to one decimal place using a TI-83+ or TI-84 calculator:

   i. The sample mean
   ii. The sample standard deviation
   iii. The median
   iv. The first quartile
   v. The third quartile
   vi. *IQR*

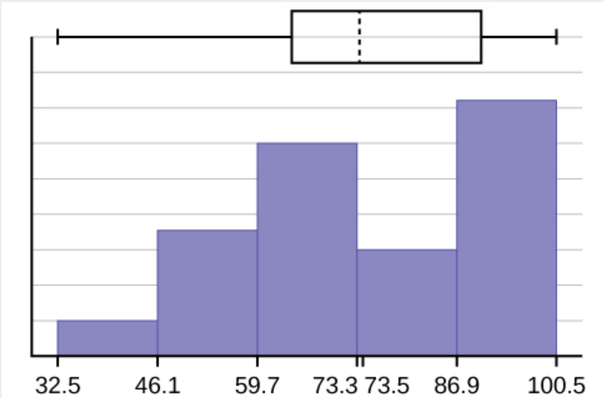
c. Construct a box plot and a histogram on the same set of axes. Make comments about the box plot, the histogram, and the chart.

Solution:

a. See [link]

b. i. The sample mean = 73.5
   ii. The sample standard deviation = 17.9
   iii. The median = 73
   iv. The first quartile = 61
   v. The third quartile = 90
   vi. *IQR* = 90 – 61 = 29

c. The *x*-axis goes from 32.5 to 100.5; *y*-axis goes from –2.4 to 15 for the histogram. The number of intervals is five, so the width of an interval is (100.5 – 32.5) divided by five, is equal to 13.6. Endpoints of the intervals are as follows: the starting point is 32.5, 32.5 + 13.6 = 46.1, 46.1 + 13.6 = 59.7, 59.7 + 13.6 = 73.3, 73.3 + 13.6 = 86.9, 86.9 + 13.6 = 100.5 = the ending value; No data values fall on an interval boundary.



The long left whisker in the box plot is reflected in the left side of the histogram. The spread of the exam scores in the lower 50% is greater (73 – 33 = 40) than the spread in the upper 50% (100 – 73 = 27). The histogram, box plot, and chart all reflect this. There are a substantial number of A and B grades (80s, 90s, and 100). The histogram clearly shows this. The box plot shows us that the middle 50% of the exam scores (*IQR* = 29) are Ds, Cs, and Bs. The box plot also shows us that the lower 25% of the exam scores are Ds and Fs.

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
33	1	0.032	0.032
42	1	0.032	0.064
49	2	0.065	0.129
53	1	0.032	0.161
55	2	0.065	0.226

Data	Frequency	Relative Frequency	Cumulative Relative Frequency
61	1	0.032	0.258
63	1	0.032	0.29
67	1	0.032	0.322
68	2	0.065	0.387
69	2	0.065	0.452
72	1	0.032	0.484
73	1	0.032	0.516
74	1	0.032	0.548
78	1	0.032	0.580
80	1	0.032	0.612
83	1	0.032	0.644
88	3	0.097	0.741
90	1	0.032	0.773
92	1	0.032	0.805
94	4	0.129	0.934
96	1	0.032	0.966
100	1	0.032	0.998 (Why isn't this value 1?)

Note:
Try It
Exercise:

Problem:

The following data show the different types of pet food stores in the area carry.
6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;
Calculate the sample mean and the sample standard deviation to one decimal place using a TI-83+ or TI-84 calculator.

Solution:

$\mu = 9.3$

$s = 2.2$

Standard deviation of Grouped Frequency Tables

Recall that for grouped data we do not know individual data values, so we cannot describe the typical value of the data with precision. In other words, we cannot find the exact mean, median, or mode. We can, however, determine the best estimate of the measures of center by finding the mean of the grouped data with the formula:

$$Mean\ of\ Frequency\ Table = \frac{\sum fm}{\sum f}$$

where f = interval frequencies and m = interval midpoints.

Just as we could not find the exact mean, neither can we find the exact standard deviation. Remember that standard deviation describes numerically the expected deviation a data value has from the mean. In simple English, the standard deviation allows us to compare how "unusual" individual data is compared to the mean.

Example:
Find the standard deviation for the data in [link].

Class	Frequency, f	Midpoint, m	m^2	x^2	fm^2	Standard Deviation

Class	Frequency, *f*	Midpoint, *m*	m^2	x^2	fm^2	Standard Deviation
0–2	1	1	1	7.58	1	3.5
3–5	6	4	16	7.58	96	3.5
6–8	10	7	49	7.58	490	3.5
9–11	7	10	100	7.58	700	3.5
12–14	0	13	169	7.58	0	3.5
15–17	2	16	256	7.58	512	3.5

For this data set, we have the mean, x = 7.58 and the standard deviation, s_x = 3.5. This means that a randomly selected data value would be expected to be 3.5 units from the mean. If we look at the first class, we see that the class midpoint is equal to one. This is almost two full standard deviations from the mean since 7.58 – 3.5 – 3.5 = 0.58. While the formula for calculating the standard deviation is not complicated, $s_x = \sqrt{\frac{f(m-x)^2}{n-1}}$ where s_x = sample standard deviation, x = sample mean, the calculations are tedious. It is usually best to use technology when performing the calculations.

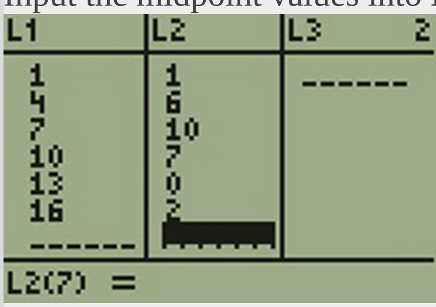
Note:
Try It
Find the standard deviation for the data from the previous example

Class	Frequency, *f*
0–2	1

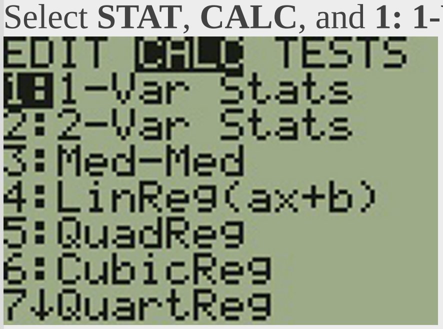
Class	Frequency, *f*
3–5	6
6–8	10
9–11	7
12–14	0
15–17	2

First, press the **STAT** key and select **1:Edit**

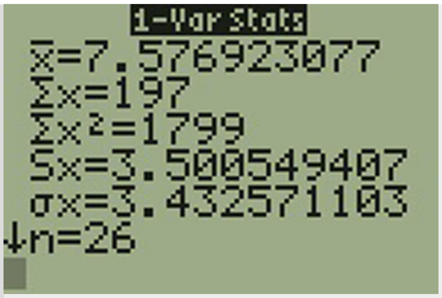
Input the midpoint values into **L1** and the frequencies into **L2**



Select **STAT**, **CALC**, and **1: 1-Var Stats**



Select **2nd** then **1** then , **2nd** then **2 Enter**



You will see displayed both a population standard deviation, σ_x, and the sample standard deviation, s_x.

Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- $\#ofSTDEVs = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$
- Compare the results of this calculation.

#ofSTDEVs is often called a "z-score"; we can use the symbol *z*. In symbols, the formulas become:

Sample	$x = x + zs$	$z = \frac{x - x}{s}$
Population	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Example:
Exercise:

Problem:

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

Solution:

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

$z = \#\text{ of STDEVs} = \frac{\text{value} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$

For John, $z = \#ofSTDEVs = \frac{2.85 - 3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77 - 80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Note:

Try It
Exercise:

Problem:

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	Time (seconds)	Team Mean Time	Team Standard Deviation
Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

Solution:

For Angie: $z = \frac{26.2 - 27.2}{0.8} = -1.25$

For Beth: $z = \frac{27.3 - 30.1}{1.4} = -2$

The following lists give a few facts that provide a little more insight into what the standard deviation tells us about the distribution of the data.

For ANY data set, no matter what the distribution of the data is:

- At least 75% of the data is within two standard deviations of the mean.
- At least 89% of the data is within three standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.
- This is known as Chebyshev's Rule.

For data having a distribution that is BELL-SHAPED and SYMMETRIC:

- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.

- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

References

Data from Microsoft Bookshelf.

King, Bill.“Graphically Speaking.” Institutional Research, Lake Tahoe Community College. Available online at http://www.ltcc.edu/web/about/institutional-research (accessed April 3, 2013).

Chapter Review

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- $s = \sqrt{\frac{\sum (x-x)^2}{n-1}}$ or $s = \sqrt{\frac{\sum f(x-x)^2}{n-1}}$ is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula $\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$ or $\sigma = \sqrt{\frac{\sum f(x-\mu)^2}{N}}$.

Formula Review

$s_x = \sqrt{\frac{\sum fm^2}{n} - x^2}$ where $\begin{array}{l} s_x = \text{ sample standard deviation} \\ x = \text{sample mean} \end{array}$

Use the following information to answer the next two exercises: The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150
Exercise:

Problem:

Use a graphing calculator or computer to find the standard deviation and round to the nearest tenth.

Solution:

s = 34.5

Exercise:

Problem: Find the value that is one standard deviation below the mean.

Exercise:

Problem:

Two baseball players, Fredo and Karl, on different teams wanted to find out who had the higher batting average when compared to his team. Which baseball player had the higher batting average when compared to his team?

Baseball Player	Batting Average	Team Batting Average	Team Standard Deviation
Fredo	0.158	0.166	0.012
Karl	0.177	0.189	0.015

Solution:

For Fredo: $z = \frac{0.158 - 0.166}{0.012} = -0.67$

For Karl: $z = \frac{0.177 - 0.189}{0.015} = -0.8$

Fredo's *z*-score of –0.67 is higher than Karl's *z*-score of –0.8. For batting average, higher values are better, so Fredo has a better batting average compared to his team.

Exercise:

Problem: Use [link] to find the value that is three standard deviations:

- **a**above the mean
- **b**below the mean

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

Exercise:

Problem:

Find the standard deviation for the following frequency tables using the formula. Check the calculations with the TI 83/84.

a.

Grade	Frequency
49.5–59.5	2
59.5–69.5	3
69.5–79.5	8
79.5–89.5	12
89.5–99.5	5

b.

Daily Low Temperature	Frequency
49.5–59.5	53

Daily Low Temperature	Frequency
59.5–69.5	32
69.5–79.5	15
79.5–89.5	1
89.5–99.5	0

c.

Points per Game	Frequency
49.5–59.5	14
59.5–69.5	32
69.5–79.5	15
79.5–89.5	23
89.5–99.5	2

Solution:

a. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{193157.45}{30} - 79.5^2} = 10.88$

b. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{380945.3}{101} - 60.94^2} = 7.62$

c. $s_x = \sqrt{\frac{\sum fm^2}{n} - x^2} = \sqrt{\frac{440051.5}{86} - 70.66^2} = 11.14$

Homework

Use the following information to answer the next nine exercises: The population parameters below describe the full-time equivalent number of students (FTES) each year at Lake Tahoe Community College from 1976–1977 through 2004–2005.

- μ = 1000 FTES
- median = 1,014 FTES
- σ = 474 FTES
- first quartile = 528.5 FTES
- third quartile = 1,447.5 FTES
- n = 29 years

Exercise:

Problem:

A sample of 11 years is taken. About how many are expected to have a FTES of 1014 or above? Explain how you determined your answer.

Solution:

The median value is the middle value in the ordered list of data values. The median value of a set of 11 will be the 6th number in order. Six years will have totals at or below the median.

Exercise:

Problem: 75% of all years have an FTES:

a. at or below: _____
b. at or above: _____

Exercise:

Problem: The population standard deviation = _____

Solution:

474 FTES

Exercise:

Problem:

What percent of the FTES were from 528.5 to 1447.5? How do you know?

Exercise:

Problem: What is the *IQR*? What does the *IQR* represent?

Solution:

919

Exercise:

Problem: How many standard deviations away from the mean is the median?

Additional Information: The population FTES for 2005–2006 through 2010–2011 was given in an updated report. The data are reported here.

Year	2005–06	2006–07	2007–08	2008–09	2009–10	2010–11
Total FTES	1,585	1,690	1,735	1,935	2,021	1,890

Exercise:

Problem:

Calculate the mean, median, standard deviation, the first quartile, the third quartile and the *IQR*. Round to one decimal place.

Solution:

- mean = 1,809.3
- median = 1,812.5
- standard deviation = 151.2
- first quartile = 1,690
- third quartile = 1,935
- *IQR* = 245

Exercise:

Problem:

What additional information is needed to construct a box plot for the FTES for 2005-2006 through 2010-2011 and a box plot for the FTES for 1976-1977 through 2004-2005?

Exercise:

Problem:

Compare the *IQR* for the FTES for 1976–77 through 2004–2005 with the *IQR* for the FTES for 2005-2006 through 2010–2011. Why do you suppose the *IQR*s are so different?

Solution:

Hint: Think about the number of years covered by each time period and what happened to higher education during those periods.

Exercise:

Problem:

Three students were applying to the same graduate school. They came from schools with different grading systems. Which student had the best GPA when compared to other students at his school? Explain how you determined your answer.

Student	GPA	School Average GPA	School Standard Deviation
Thuy	2.7	3.2	0.8
Vichet	87	75	20
Kamala	8.6	8	0.4

Exercise:

Problem:

A music school has budgeted to purchase three musical instruments. They plan to purchase a piano costing $3,000, a guitar costing $550, and a drum set costing $600. The mean cost for a piano is $4,000 with a standard deviation of $2,500. The mean cost for a guitar is $500 with a standard deviation of $200. The mean cost for drums is $700 with a standard deviation of $100. Which cost is the lowest, when compared to other instruments of the same type? Which cost is the highest when compared to other instruments of the same type. Justify your answer.

Solution:

For pianos, the cost of the piano is 0.4 standard deviations BELOW the mean. For guitars, the cost of the guitar is 0.25 standard deviations ABOVE the mean. For drums, the cost of the drum set is 1.0 standard deviations BELOW the mean. Of the three, the drums cost the lowest in comparison to the cost of other instruments of the same type. The guitar costs the most in comparison to the cost of other instruments of the same type.

Exercise:

Problem:

An elementary school class ran one mile with a mean of 11 minutes and a standard deviation of three minutes. Rachel, a student in the class, ran one mile in eight minutes. A junior high school class ran one mile with a mean of nine minutes and a standard deviation of two minutes. Kenji, a student in the class, ran 1 mile in 8.5 minutes. A high school class ran one mile with a mean of seven minutes and a standard deviation of four minutes. Nedda, a student in the class, ran one mile in eight minutes.

a. Why is Kenji considered a better runner than Nedda, even though Nedda ran faster than he?
b. Who is the fastest runner with respect to his or her class? Explain why.

Exercise:

Problem:

The most obese countries in the world have obesity rates that range from 11.4% to 74.6%. This data is summarized in Table 14.

Percent of Population Obese	Number of Countries
11.4–20.45	29
20.45–29.45	13
29.45–38.45	4
38.45–47.45	0
47.45–56.45	2
56.45–65.45	1
65.45–74.45	0
74.45–83.45	1

What is the best estimate of the average obesity percentage for these countries? What is the standard deviation for the listed obesity rates? The United States has an average obesity rate of 33.9%. Is this rate above average or below? How "unusual" is the United States' obesity rate compared to the average rate? Explain.

Solution:

- $x = 23.32$
- Using the TI 83/84, we obtain a standard deviation of: $s_x = 12.95$.
- The obesity rate of the United States is 10.58% higher than the average obesity rate.
- Since the standard deviation is 12.95, we see that 23.32 + 12.95 = 36.27 is the obesity percentage that is one standard deviation from the mean. The United States obesity rate is slightly less than one standard deviation from the mean. Therefore, we can assume that the United States, while 34% obese, does not hav e an unusually high percentage of obese people.

Exercise:

Problem:

[link] gives the percent of children under five considered to be underweight.

Percent of Underweight Children	Number of Countries
16–21.45	23
21.45–26.9	4
26.9–32.35	9
32.35–37.8	7
37.8–43.25	6
43.25–48.7	1

What is the best estimate for the mean percentage of underweight children? What is the standard deviation? Which interval(s) could be considered unusual? Explain.

Bringing It Together

Exercise:

Problem:

Twenty-five randomly selected students were asked the number of movies they watched the previous week. The results are as follows:

# of movies	Frequency
0	5
1	9
2	6
3	4

# of movies	Frequency
4	1

a. Find the sample mean x.
b. Find the approximate sample standard deviation, s.

Solution:

a. 1.48
b. 1.12

Exercise:

Problem:

Forty randomly selected students were asked the number of pairs of sneakers they owned. Let X = the number of pairs of sneakers owned. The results are as follows:

X	**Frequency**
1	2
2	5
3	8
4	12
5	12
6	0
7	1

a. Find the sample mean x

b. Find the sample standard deviation, s
c. Construct a histogram of the data.
d. Complete the columns of the chart.
e. Find the first quartile.
f. Find the median.
g. Find the third quartile.
h. Construct a box plot of the data.
i. What percent of the students owned at least five pairs?
j. Find the 40th percentile.
k. Find the 90th percentile.
l. Construct a line graph of the data
m. Construct a stemplot of the data

Exercise:

Problem:

Following are the published weights (in pounds) of all of the team members of the San Francisco 49ers from a previous year.

177; 205; 210; 210; 232; 205; 185; 185; 178; 210; 206; 212; 184; 174; 185; 242; 188; 212; 215; 247; 241; 223; 220; 260; 245; 259; 278; 270; 280; 295; 275; 285; 290; 272; 273; 280; 285; 286; 200; 215; 185; 230; 250; 241; 190; 260; 250; 302; 265; 290; 276; 228; 265

a. Organize the data from smallest to largest value.
b. Find the median.
c. Find the first quartile.
d. Find the third quartile.
e. Construct a box plot of the data.
f. The middle 50% of the weights are from _______ to _______.
g. If our population were all professional football players, would the above data be a sample of weights or the population of weights? Why?
h. If our population included every team member who ever played for the San Francisco 49ers, would the above data be a sample of weights or the population of weights? Why?
i. Assume the population was the San Francisco 49ers. Find:

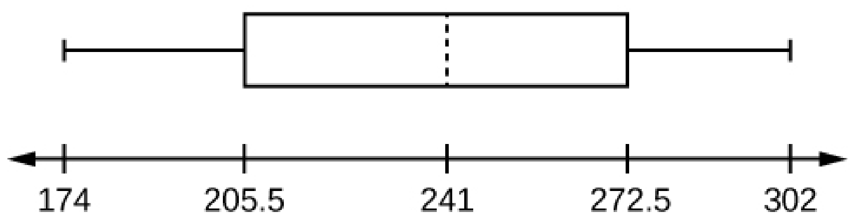
    i. the population mean, μ.
    ii. the population standard deviation, σ.
    iii. the weight that is two standard deviations below the mean.
    iv. When Steve Young, quarterback, played football, he weighed 205 pounds. How many standard deviations above or below the mean was he?

j. That same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds. Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young? How did you determine your answer?

Solution:

a. 174; 177; 178; 184; 185; 185; 185; 185; 188; 190; 200; 205; 205; 206; 210; 210; 210; 212; 212; 215; 215; 220; 223; 228; 230; 232; 241; 241; 242; 245; 247; 250; 250; 259; 260; 260; 265; 265; 270; 272; 273; 275; 276; 278; 280; 280; 285; 285; 286; 290; 290; 295; 302
b. 241
c. 205.5
d. 272.5

e.



f. 205.5, 272.5
g. sample
h. population

i.
  i. 236.34
  ii. 37.50
  iii. 161.34
  iv. 0.84 std. dev. below the mean

j. Young

Exercise:

Problem:

One hundred teachers attended a seminar on mathematical problem solving. The attitudes of a representative sample of 12 of the teachers were measured before and after the seminar. A positive number for change in attitude indicates that a teacher's attitude toward math became more positive. The 12 change scores are as follows:

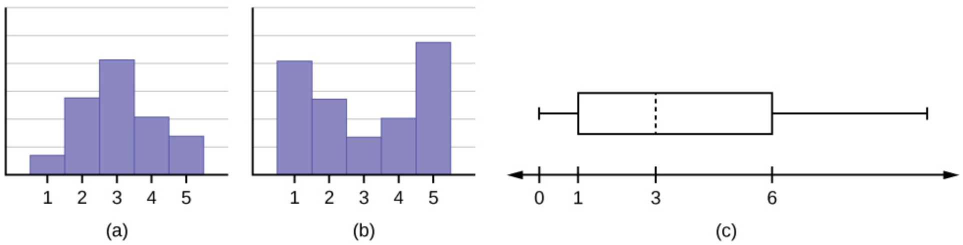
3 8–12 05–31–16 5–2

a. What is the mean change score?
b. What is the standard deviation for this population?
c. What is the median change score?
d. Find the change score that is 2.2 standard deviations below the mean.

Exercise:

Problem:

Refer to [link] determine which of the following are true and which are false. Explain your solution to each part in complete sentences.



a. The medians for all three graphs are the same.
b. We cannot determine if any of the means for the three graphs is different.
c. The standard deviation for graph b is larger than the standard deviation for graph a.
d. We cannot determine if any of the third quartiles for the three graphs is different.

Solution:

a. True
b. True
c. True
d. False

Exercise:

Problem:

In a recent issue of the *IEEE Spectrum*, 84 engineering conferences were announced. Four conferences lasted two days. Thirty-six lasted three days. Eighteen lasted four days. Nineteen lasted five days. Four lasted six days. One lasted seven days. One lasted eight days. One lasted nine days. Let X = the length (in days) of an engineering conference.

a. Organize the data in a chart.

b. Find the median, the first quartile, and the third quartile.
c. Find the 65th percentile.
d. Find the 10th percentile.
e. Construct a box plot of the data.
f. The middle 50% of the conferences last from _______ days to _______ days.
g. Calculate the sample mean of days of engineering conferences.
h. Calculate the sample standard deviation of days of engineering conferences.
i. Find the mode.
j. If you were planning an engineering conference, which would you choose as the length of the conference: mean; median; or mode? Explain why you made that choice.
k. Give two reasons why you think that three to five days seem to be popular lengths of engineering conferences.

Exercise:

Problem:

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

a. Organize the data into a chart with five intervals of equal width. Label the two columns "Enrollment" and "Frequency."
b. Construct a histogram of the data.
c. If you were to build a new community college, which piece of information would be more valuable: the mode or the mean?
d. Calculate the sample mean.
e. Calculate the sample standard deviation.
f. A school with an enrollment of 8000 would be how many standard deviations away from the mean?

Solution:

a.

Enrollment	Frequency
1000-5000	10
5000-10000	16
10000-15000	3
15000-20000	3
20000-25000	1
25000-30000	2

b. Check student's solution.
c. mode
d. 8628.74
e. 6943.88
f. –0.09

Use the following information to answer the next two exercises. X = the number of days per week that 100 clients use a particular exercise facility.

x	Frequency
0	3
1	12
2	33
3	28
4	11

x	**Frequency**
5	9
6	4

Exercise:

Problem: The 80th percentile is _____

a. 5
b. 80
c. 3
d. 4

Exercise:

Problem:

The number that is 1.5 standard deviations BELOW the mean is approximately _____

a. 0.7
b. 4.8
c. –2.8
d. Cannot be determined

Solution:

a

Exercise:

Problem:

Suppose that a publisher conducted a survey asking adult consumers the number of fiction paperback books they had purchased in the previous month. The results are summarized in the [link].

# of books	Freq.	Rel. Freq.
0	18	
1	24	
2	24	
3	22	
4	15	
5	10	
7	5	
9	1	

a. Are there any outliers in the data? Use an appropriate numerical test involving the *IQR* to identify outliers, if any, and clearly state your conclusion.
b. If a data value is identified as an outlier, what should be done about it?
c. Are any data values further than two standard deviations away from the mean? In some situations, statisticians may use this criteria to identify data values that are unusual, compared to the other data values. (Note that this criteria is most appropriate to use for data that is mound-shaped and symmetric, rather than for skewed data.)
d. Do parts a and c of this problem give the same answer?
e. Examine the shape of the data. Which part, a or c, of this question gives a more appropriate result for this data?
f. Based on the shape of the data which is the most appropriate measure of center for this data: mean, median or mode?

Glossary

Standard Deviation
: a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: s for sample standard deviation and σ for population standard deviation.

Variance

mean of the squared deviations from the mean, or the square of the standard deviation; for a set of data, a deviation can be represented as $x - x$ where x is a value of the data and x is the sample mean. The sample variance is equal to the sum of the squares of the deviations divided by the difference of the sample size and one.

Review Exercises (Ch 3-13)

These review exercises are designed to provide extra practice on concepts learned before a particular chapter. For example, the review exercises for Chapter 3, cover material learned in chapters 1 and 2.

Chapter 3

Use the following information to answer the next six exercises: In a survey of 100 stocks on NASDAQ, the average percent increase for the past year was 9% for NASDAQ stocks.

1. The "average increase" for all NASDAQ stocks is the:

a. population
b. statistic
c. parameter
d. sample
e. variable

2. All of the NASDAQ stocks are the:

a. population
b. statistics
c. parameter
d. sample
e. variable

3. Nine percent is the:

a. population
b. statistics
c. parameter
d. sample
e. variable

4. The 100 NASDAQ stocks in the survey are the:

a. population
b. statistic
c. parameter
d. sample
e. variable

5. The percent increase for one stock in the survey is the:

a. population
b. statistic
c. parameter
d. sample
e. variable

6. Would the data collected by qualitative, quantitative discrete, or quantitative continuous?

Use the following information to answer the next two exercises: Thirty people spent two weeks around Mardi Gras in New Orleans. Their two-week weight gain is below. (Note: a loss is shown by a negative weight gain.)

Weight Gain	**Frequency**
–2	3
–1	5
0	2
1	4
4	13
6	2

Weight Gain	Frequency
11	1

7. Calculate the following values:

a. the average weight gain for the two weeks
b. the standard deviation
c. the first, second, and third quartiles

8. Construct a histogram and box plot of the data.

Chapter 4

Use the following information to answer the next two exercises: A recent poll concerning credit cards found that 35 percent of respondents use a credit card that gives them a mile of air travel for every dollar they charge. Thirty percent of the respondents charge more than $2,000 per month. Of those respondents who charge more than $2,000, 80 percent use a credit card that gives them a mile of air travel for every dollar they charge.

9. What is the probability that a randomly selected respondent will spend more than $2,000 AND use a credit card that gives them a mile of air travel for every dollar they charge?

a. (0.30)(0.35)
b. (0.80)(0.35)
c. (0.80)(0.30)
d. (0.80)

10. Are using a credit card that gives a mile of air travel for each dollar spent AND charging more than $2,000 per month independent events?

a. Yes
b. No, and they are not mutually exclusive either.
c. No, but they are mutually exclusive.
d. Not enough information given to determine the answer

11. A sociologist wants to know the opinions of employed adult women about government funding for day care. She obtains a list of 520 members of a local business and professional women's club and mails a questionnaire to 100 of these women selected at random. Sixty-eight questionnaires are returned. What is the population in this study?

a. all employed adult women
b. all the members of a local business and professional women's club
c. the 100 women who received the questionnaire
d. all employed women with children

Use the following information to answer the next two exercises: The next two questions refer to the following: An article from The San Jose Mercury News was concerned with the racial mix of the 1500 students at Prospect High School in Saratoga, CA. The table summarizes the results. (Male and female values are approximate.) Suppose one Prospect High School student is randomly selected.

Gender/Ethnic group	White	Asian	Hispanic	Black	American Indian
Male	400	468	115	35	16
Female	440	132	140	40	14

12. Find the probability that a student is Asian or Male.

13. Find the probability that a student is Black given that the student is female.

14. A sample of pounds lost, in a certain month, by individual members of a weight reducing clinic produced the following statistics:

- Mean = 5 lbs.
- Median = 4.5 lbs.

- Mode = 4 lbs.
- Standard deviation = 3.8 lbs.
- First quartile = 2 lbs.
- Third quartile = 8.5 lbs.

The correct statement is:

a. One fourth of the members lost exactly two pounds.
b. The middle fifty percent of the members lost from two to 8.5 lbs.
c. Most people lost 3.5 to 4.5 lbs.
d. All of the choices above are correct.

15. What does it mean when a data set has a standard deviation equal to zero?

a. All values of the data appear with the same frequency.
b. The mean of the data is also zero.
c. All of the data have the same value.
d. There are no data to begin with.

16. The statement that describe the illustration is:

a. the mean is equal to the median.
b. There is no first quartile.
c. The lowest data value is the median.
d. The median equals $\frac{Q_1+Q_3}{2}$.

17. According to a recent article in the *San Jose Mercury News* the average number of babies born with significant hearing loss (deafness) is approximately 2 per 1000 babies in a healthy baby nursery. The number climbs to an average of 30 per 1000 babies in an intensive care nursery. Suppose that 1,000 babies from healthy baby

nurseries were randomly surveyed. Find the probability that exactly two babies were born deaf.

18. A “friend” offers you the following “deal.” For a $10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.

- Ten of the coupons are for a free gift worth $6.
- Eighty of the coupons are for a free gift worth $8.
- Six of the coupons are for a free gift worth $12.
- Four of the coupons are for a free gift worth $40.

Based upon the financial gain or loss over the long run, should you play the game?

a. Yes, I expect to come out ahead in money.
b. No, I expect to come out behind in money.
c. It doesn’t matter. I expect to break even.

Use the following information to answer the next four exercises: Recently, a nurse commented that when a patient calls the medical advice line claiming to have the flu, the chance that he/she truly has the flu (and not just a nasty cold) is only about 4%. Of the next 25 patients calling in claiming to have the flu, we are interested in how many actually have the flu.

19. Define the random variable and list its possible values.

20. State the distribution of X.

21. Find the probability that at least four of the 25 patients actually have the flu.

22. On average, for every 25 patients calling in, how many do you expect to have the flu?

Use the following information to answer the next two exercises: Different types of writing can sometimes be distinguished by the number of letters in the words used. A student interested in this fact wants to study the number of letters of words used by Tom Clancy in his novels. She opens a Clancy novel at random and records the number of letters of the first 250 words on the page.

23. What kind of data was collected?

a. qualitative
b. quantitative continuous
c. quantitative discrete

24. What is the population under study?

Chapter 5

Use the following information to answer the next seven exercises: A recent study of mothers of junior high school children in Santa Clara County reported that 76% of the mothers are employed in paid positions. Of those mothers who are employed, 64% work full-time (over 35 hours per week), and 36% work part-time. However, out of all of the mothers in the population, 49% work full-time. The population under study is made up of mothers of junior high school children in Santa Clara County. Let E = employed and F = full-time employment.

25.

a. Find the percent of all mothers in the population that are NOT employed.
b. Find the percent of mothers in the population that are employed part-time.

26. The "type of employment" is considered to be what type of data?

27. Find the probability that a randomly selected mother works part-time given that she is employed.

28. Find the probability that a randomly selected person from the population will be employed or work full-time.

29. Being employed and working part-time:

a. mutually exclusive events? Why or why not?
b. independent events? Why or why not?

Use the following additional information to answer the next two exercises: We randomly pick ten mothers from the above population. We are interested in the

number of the mothers that are employed. Let X = number of mothers that are employed.

30. State the distribution for X.

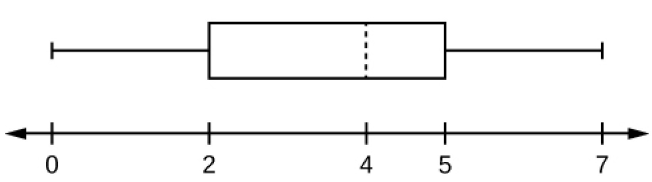
31. Find the probability that at least six are employed.

32. We expect the statistics discussion board to have, on average, 14 questions posted to it per week. We are interested in the number of questions posted to it per day.

a. Define X.
b. What are the values that the random variable may take on?
c. State the distribution for X.
d. Find the probability that from ten to 14 (inclusive) questions are posted to the listserv on a randomly picked day.

33. A person invests $1,000 into stock of a company that hopes to go public in one year. The probability that the person will lose all his money after one year (i.e. his stock will be worthless) is 35%. The probability that the person's stock will still have a value of $1,000 after one year (i.e. no profit and no loss) is 60%. The probability that the person's stock will increase in value by $10,000 after one year (i.e. will be worth $11,000) is 5%. Find the expected profit after one year.

34. Rachel's piano cost $3,000. The average cost for a piano is $4,000 with a standard deviation of $2,500. Becca's guitar cost $550. The average cost for a guitar is $500 with a standard deviation of $200. Matt's drums cost $600. The average cost for drums is $700 with a standard deviation of $100. Whose cost was lowest when compared to his or her own instrument?

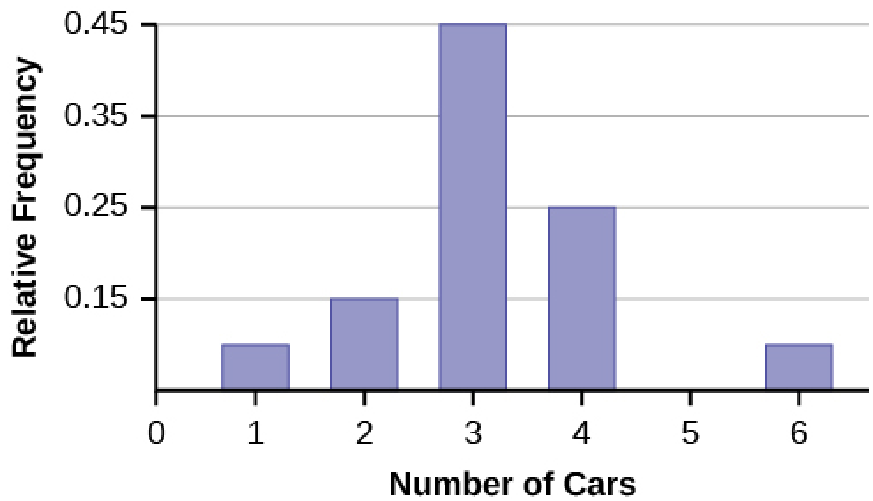


35. Explain why each statement is either true or false given the box plot in [link].

a. Twenty-five percent of the data re at most five.
b. There is the same amount of data from 4–5 as there is from 5–7.
c. There are no data values of three.

d. Fifty percent of the data are four.

Using the following information to answer the next two exercises: 64 faculty members were asked the number of cars they owned (including spouse and children's cars). The results are given in the following graph:



36. Find the approximate number of responses that were three.

37. Find the first, second and third quartiles. Use them to construct a box plot of the data.

Use the following information to answer the next three exercises: [link] shows data gathered from 15 girls on the Snow Leopard soccer team when they were asked how they liked to wear their hair. Supposed one girl from the team is randomly selected.

Hair Style/Hair Color	**Blond**	**Brown**	**Black**
Ponytail	3	2	5
Plain	2	2	1

38. Find the probability that the girl has black hair GIVEN that she wears a ponytail.

39. Find the probability that the girl wears her hair plain OR has brown hair.

40. Find the probability that the girl has blond hair AND that she wears her hair plain.

Chapter 6

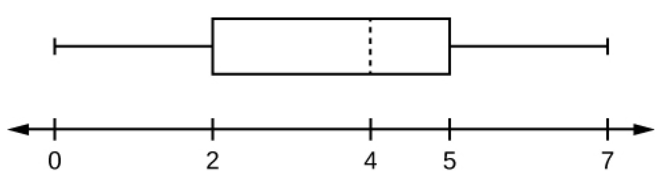
Use the following information to answer the next two exercises: $X \sim U(3, 13)$

41. Explain which of the following are false and which are true.

a. $f(x) = \frac{1}{10}$, $3 \le x \le 13$
b. There is no mode
c. The median is less than the mean.
d. $P(x > 10) = P(x \le 6)$

42. Calculate:

a. the mean.
b. the median.
c. the 65th percentile.



43. Which of the following is true for the box plot in [link]?

a. Twenty-five percent of the data are at most five.
b. There is about the same amount of data from 4–5 as there is from 5–7.
c. There are no data values of three.
d. Fifty percent of the data are four.

44. If $P(G|H) = P(G)$, then which of the following is correct?

a. G and H are mutually exclusive events.
b. $P(G) = P(H)$
c. Knowing that H has occurred will affect the chance that G will happen.
d. G and H are independent events.

45. If $P(J) = 0.3$, $P(K) = 0.63$, and J and K are independent events, then explain which are correct and which are incorrect.

a. $P(J \text{ AND } K) = 0$
b. $P(J \text{ OR } K) = 0.9$
c. $P(J \text{ OR } K) = 0.72$
d. $P(J) \neq P(J|K)$

46. On average, five students from each high school class get full scholarships to four-year colleges. Assume that most high school classes have about 500 students. X = the number of students from a high school class that get full scholarships to four-year schools. Which of the following is the distribution of X?

a. $P(5)$
b. $B(500, 5)$
c. $Exp\left(\frac{1}{5}\right)$
d. $N\left(5, \frac{(0.01)(0.99)}{500}\right)$

Chapter 7

Use the following information to answer the next three exercises: Richard's Furniture Company delivers furniture from 10 A.M. to 2 P.M. continuously and uniformly. We are interested in how long (in hours) past the 10 A.M. start time that individuals wait for their delivery.

47. X ~ __________

a. $U(0, 4)$
b. $U(10, 20)$

c. $Exp(2)$
d. $N(2, 1)$

48. The average wait time is:

a. 1 hour.
b. 2 hours.
c. 2.5 hours.
d. 4 hours.

49. Suppose that it is now past noon on a delivery day. The probability that a person must wait at least 1.5 more hours is:

a. $\frac{1}{4}$
b. $\frac{1}{2}$
c. $\frac{3}{4}$
d. $\frac{3}{8}$

50. Given: $X \sim Exp\left(\frac{1}{3}\right)$

a. Find $P(x > 1)$.
b. Calculate the minimum value for the upper quartile.
c. Find $P\left(x = \frac{1}{3}\right)$

51.

- 40% of full-time students took 4 years to graduate
- 30% of full-time students took 5 years to graduate
- 20% of full-time students took 6 years to graduate
- 10% of full-time students took 7 years to graduate

The expected time for full-time students to graduate is:

a. 4 years

b. 4.5 years
c. 5 years
d. 5.5 years

52. Which of the following distributions is described by the following example? Many people can run a short distance of under two miles, but as the distance increases, fewer people can run that far.

a. binomial
b. uniform
c. exponential
d. normal

53. The length of time to brush one's teeth is generally thought to be exponentially distributed with a mean of $\frac{3}{4}$ minutes. Find the probability that a randomly selected person brushes his or her teeth less than $\frac{3}{4}$ minutes.

a. 0.5
b. $\frac{3}{4}$
c. 0.43
d. 0.63

54. Which distribution accurately describes the following situation?
The chance that a teenage boy regularly gives his mother a kiss goodnight is about 20%. Fourteen teenage boys are randomly surveyed. Let X = the number of teenage boys that regularly give their mother a kiss goodnight.

a. $B(14,0.20)$
b. $P(2.8)$
c. $N(2.8,2.24)$
d. $Exp\left(\frac{1}{0.20}\right)$

55. A 2008 report on technology use states that approximately 20% of U.S. households have never sent an e-mail. Suppose that we select a random sample of

fourteen U.S. households. Let X = the number of households in a 2008 sample of 14 households that have never sent an email

a. $B(14,0.20)$
b. $P(2.8)$
c. $N(2.8,2.24)$
d. $Exp\left(\frac{1}{0.20}\right)$

Chapter 8

Use the following information to answer the next three exercises: Suppose that a sample of 15 randomly chosen people were put on a special weight loss diet. The amount of weight lost, in pounds, follows an unknown distribution with mean equal to 12 pounds and standard deviation equal to three pounds. Assume that the distribution for the weight loss is normal.

56. To find the probability that the mean amount of weight lost by 15 people is no more than 14 pounds, the random variable should be:

a. number of people who lost weight on the special weight loss diet.
b. the number of people who were on the diet.
c. the mean amount of weight lost by 15 people on the special weight loss diet.
d. the total amount of weight lost by 15 people on the special weight loss diet.

57. Find the probability asked for in Question 56.

58. Find the 90^{th} percentile for the mean amount of weight lost by 15 people.

Using the following information to answer the next three exercises: The time of occurrence of the first accident during rush-hour traffic at a major intersection is uniformly distributed between the three hour interval 4 p.m. to 7 p.m. Let X = the amount of time (hours) it takes for the first accident to occur.

59. What is the probability that the time of occurrence is within the first half-hour or the last hour of the period from 4 to 7 p.m.?

a. cannot be determined from the information given
b. $\frac{1}{6}$

c. $\frac{1}{2}$
d. $\frac{1}{3}$

60. The 20th percentile occurs after how many hours?

a. 0.20
b. 0.60
c. 0.50
d. 1

61. Assume Ramon has kept track of the times for the first accidents to occur for 40 different days. Let *C* = the total cumulative time. Then *C* follows which distribution?

a. *U*(0,3)
b. *Exp*(13)
c. *N*(60, 5.477)
d. *N*(1.5, 0.01875)

62. Using the information in Question 61, find the probability that the total time for all first accidents to occur is more than 43 hours.

Use the following information to answer the next two exercises: The length of time a parent must wait for his children to clean their rooms is uniformly distributed in the time interval from one to 15 days.

63. How long must a parent expect to wait for his children to clean their rooms?

a. eight days
b. three days
c. 14 days
d. six days

64. What is the probability that a parent will wait more than six days given that the parent has already waited more than three days?

a. 0.5174
b. 0.0174
c. 0.7500
d. 0.2143

Use the following information to answer the next five exercises: Twenty percent of the students at a local community college live in within five miles of the campus. Thirty percent of the students at the same community college receive some kind of financial aid. Of those who live within five miles of the campus, 75% receive some kind of financial aid.

65. Find the probability that a randomly chosen student at the local community college does not live within five miles of the campus.

a. 80%
b. 20%
c. 30%
d. cannot be determined

66. Find the probability that a randomly chosen student at the local community college lives within five miles of the campus or receives some kind of financial aid.

a. 50%
b. 35%
c. 27.5%
d. 75%

67. Are living in student housing within five miles of the campus and receiving some kind of financial aid mutually exclusive?

a. yes
b. no
c. cannot be determined

68. The interest rate charged on the financial aid is ________ data.

a. quantitative discrete
b. quantitative continuous
c. qualitative discrete
d. qualitative

69. The following information is about the students who receive financial aid at the local community college.

- 1st quartile = $250
- 2nd quartile = $700
- 3rd quartile = $1200

These amounts are for the school year. If a sample of 200 students is taken, how many are expected to receive $250 or more?

a. 50
b. 250
c. 150
d. cannot be determined

Use the following information to answer the next two exercises: $P(A) = 0.2$, $P(B) = 0.3$; *A* and *B* are independent events.

70. $P(A \text{ AND } B)$ = _______

a. 0.5
b. 0.6
c. 0
d. 0.06

71. $P(A \text{ OR } B)$ = ________

a. 0.56
b. 0.5
c. 0.44
d. 1

72. If H and D are mutually exclusive events, $P(H) = 0.25$, $P(D) = 0.15$, then $P(H|D)$.

a. 1
b. 0
c. 0.40
d. 0.0375

Chapter 9

73. Rebecca and Matt are 14 year old twins. Matt's height is two standard deviations below the mean for 14 year old boys' height. Rebecca's height is 0.10 standard deviations above the mean for 14 year old girls' height. Interpret this.

a. Matt is 2.1 inches shorter than Rebecca.
b. Rebecca is very tall compared to other 14 year old girls.
c. Rebecca is taller than Matt.
d. Matt is shorter than the average 14 year old boy.

74. Construct a histogram of the IPO data (see [link]).

Use the following information to answer the next three exercises: Ninety homeowners were asked the number of estimates they obtained before having their homes fumigated. Let X = the number of estimates.

x	Relative Frequency	Cumulative Relative Frequency
1	0.3	
2	0.2	
4	0.4	

x	Relative Frequency	Cumulative Relative Frequency
5	0.1	

75. Complete the cumulative frequency column.

76. Calculate the sample mean (a), the sample standard deviation (b) and the percent of the estimates that fall at or below four (c).

77. Calculate the median, *M*, the first quartile, Q_1, the third quartile, Q_3. Then construct a box plot of the data.

78. The middle 50% of the data are between _____ and _____.

Use the following information to answer the next three exercises: Seventy 5th and 6th graders were asked their favorite dinner.

	Pizza	Hamburgers	Spaghetti	Fried shrimp
5th grader	15	6	9	0
6th grader	15	7	10	8

79. Find the probability that one randomly chosen child is in the 6th grade and prefers fried shrimp.

a. $\frac{32}{70}$
b. $\frac{8}{32}$
c. $\frac{8}{8}$
d. $\frac{8}{70}$

80. Find the probability that a child does not prefer pizza.

a. $\frac{30}{70}$
b. $\frac{30}{40}$
c. $\frac{40}{70}$
d. 1

81. Find the probability a child is in the 5th grade given that the child prefers spaghetti.

a. $\frac{9}{19}$
b. $\frac{9}{70}$
c. $\frac{9}{30}$
d. $\frac{19}{70}$

82. A sample of convenience is a random sample.

a. true
b. false

83. A statistic is a number that is a property of the population.

a. true
b. false

84. You should always throw out any data that are outliers.

a. true
b. false

85. Lee bakes pies for a small restaurant in Felton, CA. She generally bakes 20 pies in a day, on average. Of interest is the number of pies she bakes each day.

a. Define the random variable X.

b. State the distribution for *X*.
c. Find the probability that Lee bakes more than 25 pies in any given day.

86. Six different brands of Italian salad dressing were randomly selected at a supermarket. The grams of fat per serving are 7, 7, 9, 6, 8, 5. Assume that the underlying distribution is normal. Calculate a 95% confidence interval for the population mean grams of fat per serving of Italian salad dressing sold in supermarkets.

87. Given: uniform, exponential, normal distributions. Match each to a statement below.

a. mean = median ≠ mode
b. mean > median > mode
c. mean = median = mode

Chapter 10

Use the following information to answer the next three exercises: In a survey at Kirkwood Ski Resort the following information was recorded:

	0–10	**11–20**	**21–40**	**40+**
Ski	10	12	30	8
Snowboard	6	17	12	5

Suppose that one person from [link] was randomly selected.

88. Find the probability that the person was a skier or was age 11–20.

89. Find the probability that the person was a snowboarder given he or she was age 21–40.

90. Explain which of the following are true and which are false.

a. Sport and age are independent events.
b. Ski and age 11–20 are mutually exclusive events.
c. P(Ski AND age 21–40) < P(Ski|age 21–40)
d. P(Snowboard OR age 0–10) < P(Snowboard|age 0–10)

91. The average length of time a person with a broken leg wears a cast is approximately six weeks. The standard deviation is about three weeks. Thirty people who had recently healed from broken legs were interviewed. State the distribution that most accurately reflects total time to heal for the thirty people.

92. The distribution for X is uniform. What can we say for certain about the distribution for X when $n = 1$?

a. The distribution for X is still uniform with the same mean and standard deviation as the distribution for X.
b. The distribution for X is normal with the different mean and a different standard deviation as the distribution for X.
c. The distribution for X is normal with the same mean but a larger standard deviation than the distribution for X.
d. The distribution for X is normal with the same mean but a smaller standard deviation than the distribution for X.

93. The distribution for X is uniform. What can we say for certain about the distribution for $\sum X$ when $n = 50$?

a. distribution for $\sum X$ is still uniform with the same mean and standard deviation as the distribution for X.
b. The distribution for $\sum X$ is normal with the same mean but a larger standard deviation as the distribution for X.
c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for X.
d. The distribution for $\sum X$ is normal with the same mean but a smaller standard deviation than the distribution for X.

Use the following information to answer the next three exercises: A group of students measured the lengths of all the carrots in a five-pound bag of baby carrots. They calculated the average length of baby carrots to be 2.0 inches with a standard deviation of 0.25 inches. Suppose we randomly survey 16 five-pound bags of baby carrots.

94. State the approximate distribution for X, the distribution for the average lengths of baby carrots in 16 five-pound bags. X ~ ______

95. Explain why we cannot find the probability that one individual randomly chosen carrot is greater than 2.25 inches.

96. Find the probability that x is between two and 2.25 inches.

Use the following information to answer the next three exercises: At the beginning of the term, the amount of time a student waits in line at the campus store is normally distributed with a mean of five minutes and a standard deviation of two minutes.

97. Find the 90th percentile of waiting time in minutes.

98. Find the median waiting time for one student.

99. Find the probability that the average waiting time for 40 students is at least 4.5 minutes.

Chapter 11

Use the following information to answer the next four exercises: Suppose that the time that owners keep their cars (purchased new) is normally distributed with a mean of seven years and a standard deviation of two years. We are interested in how long an individual keeps his car (purchased new). Our population is people who buy their cars new.

100. Sixty percent of individuals keep their cars **at most** how many years?

101. Suppose that we randomly survey one person. Find the probability that person keeps his or her car **less than** 2.5 years.

102. If we are to pick individuals ten at a time, find the distribution for the **mean** car length ownership.

103. If we are to pick ten individuals, find the probability that the **sum** of their ownership time is more than 55 years.

104. For which distribution is the median not equal to the mean?

a. Uniform
b. Exponential
c. Normal
d. Student *t*

105. Compare the standard normal distribution to the Student's *t*-distribution, centered at zero. Explain which of the following are true and which are false.

a. As the number surveyed increases, the area to the left of –1 for the Student's *t*-distribution approaches the area for the standard normal distribution.
b. As the degrees of freedom decrease, the graph of the Student's *t*-distribution looks more like the graph of the standard normal distribution.
c. If the number surveyed is 15, the normal distribution should never be used.

Use the following information to answer the next five exercises: We are interested in the checking account balance of twenty-year-old college students. We randomly survey 16 twenty-year-old college students. We obtain a sample mean of \$640 and a sample standard deviation of \$150. Let X = checking account balance of an individual twenty year old college student.

106. Explain why we cannot determine the distribution of X.

107. If you were to create a confidence interval or perform a hypothesis test for the population mean checking account balance of twenty-year-old college students, what distribution would you use?

108. Find the 95% confidence interval for the true mean checking account balance of a twenty-year-old college student.

109. What type of data is the balance of the checking account considered to be?

110. What type of data is the number of twenty-year-olds considered to be?

111. On average, a busy emergency room gets a patient with a shotgun wound about once per week. We are interested in the number of patients with a shotgun wound the emergency room gets per 28 days.

a. Define the random variable X.

b. State the distribution for *X*.
c. Find the probability that the emergency room gets no patients with shotgun wounds in the next 28 days.

Use the following information to answer the next two exercises: The probability that a certain slot machine will pay back money when a quarter is inserted is 0.30. Assume that each play of the slot machine is independent from each other. A person puts in 15 quarters for 15 plays.

112. Is the expected number of plays of the slot machine that will pay back money greater than, less than or the same as the median? Explain your answer.

113. Is it likely that exactly eight of the 15 plays would pay back money? Justify your answer numerically.

114. A game is played with the following rules:

- it costs $10 to enter.
- a fair coin is tossed four times.
- if you do not get four heads or four tails, you lose your $10.
- if you get four heads or four tails, you get back your $10, plus $30 more.

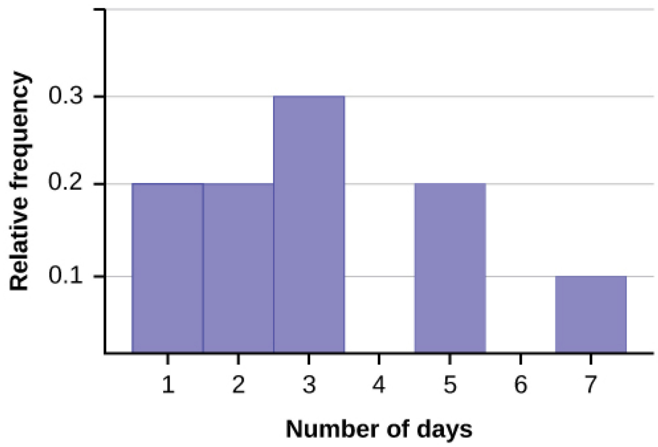
Over the long run of playing this game, what are your expected earnings?

115.

- The mean grade on a math exam in Rachel’s class was 74, with a standard deviation of five. Rachel earned an 80.
- The mean grade on a math exam in Becca’s class was 47, with a standard deviation of two. Becca earned a 51.
- The mean grade on a math exam in Matt’s class was 70, with a standard deviation of eight. Matt earned an 83.

Find whose score was the best, compared to his or her own class. Justify your answer numerically.

Use the following information to answer the next two exercises: A random sample of 70 compulsive gamblers were asked the number of days they go to casinos per week. The results are given in the following graph:



116. Find the number of responses that were five.

117. Find the mean, standard deviation, the median, the first quartile, the third quartile and the *IQR*.

118. Based upon research at De Anza College, it is believed that about 19% of the student population speaks a language other than English at home. Suppose that a study was done this year to see if that percent has decreased. Ninety-eight students were randomly surveyed with the following results. Fourteen said that they speak a language other than English at home.

a. State an appropriate null hypothesis.
b. State an appropriate alternative hypothesis.
c. Define the random variable, P'.
d. Calculate the test statistic.
e. Calculate the p-value.
f. At the 5% level of decision, what is your decision about the null hypothesis?
g. What is the Type I error?
h. What is the Type II error?

119. Assume that you are an emergency paramedic called in to rescue victims of an accident. You need to help a patient who is bleeding profusely. The patient is also considered to be a high risk for contracting AIDS. Assume that the null hypothesis is that the patient does **not** have the HIV virus. What is a Type I error?

120. It is often said that Californians are more casual than the rest of Americans. Suppose that a survey was done to see if the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals. Fifty of each was surveyed with the following results. Fifteen Californians wear jeans

to work and six non-Californians wear jeans to work.
Let *C* = Californian professional; *NC* = non-Californian professional

a. State appropriate null and alternate hypotheses.
b. Define the random variable.
c. Calculate the test statistic and *p*-value.
d. At the 5% significance level, what is your decision?
e. What is the Type I error?
f. What is the Type II error?

Use the following information to answer the next two exercises: A group of Statistics students have developed a technique that they feel will lower their anxiety level on statistics exams. They measured their anxiety level at the start of the quarter and again at the end of the quarter. Recorded is the paired data in that order: (1000, 900); (1200, 1050); (600, 700); (1300, 1100); (1000, 900); (900, 900).

121. This is a test of (pick the best answer):

a. large samples, independent means
b. small samples, independent means
c. dependent means

122. State the distribution to use for the test.

Chapter 12

Use the following information to answer the next two exercises: A recent survey of U.S. teenage pregnancy was answered by 720 girls, age 12–19. Six percent of the girls surveyed said they have been pregnant. We are interested in the true proportion of U.S. girls, age 12–19, who have been pregnant.

123. Find the 95% confidence interval for the true proportion of U.S. girls, age 12–19, who have been pregnant.

124. The report also stated that the results of the survey are accurate to within ±3.7% at the 95% confidence level. Suppose that a new study is to be done. It is desired to be accurate to within 2% of the 95% confidence level. What is the minimum number that should be surveyed?

125. Given: $X \sim Exp\left(\frac{1}{3}\right)$. Sketch the graph that depicts: $P(x > 1)$.

Use the following information to answer the next three exercises: The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the mean amount of money a customer spends in one trip to the supermarket is $72.

126. Find the probability that one customer spends less than $72 in one trip to the supermarket?

127. Suppose five customers pool their money. How much money altogether would you expect the five customers to spend in one trip to the supermarket (in dollars)?

128. State the distribution to use if you want to find the probability that the **mean** amount spent by five customers in one trip to the supermarket is less than $60.

Chapter 13

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is 10%. However, in any year, the probability of water rationing is 5%.

129. What is the probability of both a drought **and** water rationing occurring?

130. Out of the years with water rationing, find the probability that there is a drought.

Use the following information to answer the next three exercises:

	Apple	**Pumpkin**	**Pecan**
Female	40	10	30
Male	20	30	10

131. Suppose that one individual is randomly chosen. Find the probability that the person’s favorite pie is apple **or** the person is male.

132. Suppose that one male is randomly chosen. Find the probability his favorite pie is pecan.

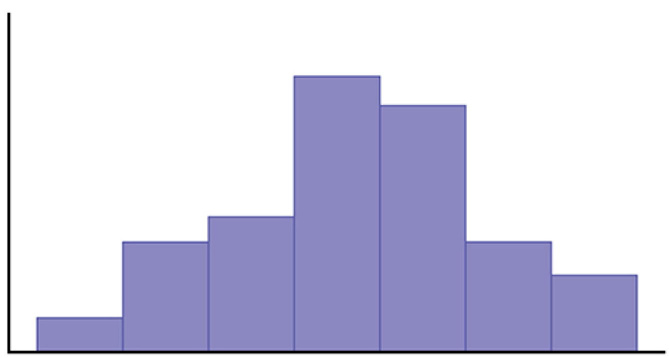
133. Conduct a hypothesis test to determine if favorite pie type and gender are independent.

Use the following information to answer the next two exercises: Let’s say that the probability that an adult watches the news at least once per week is 0.60.

134. We randomly survey 14 people. On average, how many people do we expect to watch the news at least once per week?

135. We randomly survey 14 people. Of interest is the number that watch the news at least once per week. State the distribution of X. $X \sim$ ______

136. The following histogram is most likely to be a result of sampling from which distribution?



a. Chi-Square
b. Geometric
c. Uniform
d. Binomial

137. The ages of De Anza evening students is known to be normally distributed with a population mean of 40 and a population standard deviation of six. A sample of six De Anza evening students reported their ages (in years) as: 28; 35; 47; 45; 30; 50. Find the probability that the mean of six ages of randomly chosen students is less than 35 years. Hint: Find the sample mean.

138. A math exam was given to all the fifth grade children attending Country School. Two random samples of scores were taken. The null hypothesis is that the mean math scores for boys and girls in fifth grade are the same. Conduct a hypothesis test.

	n	x	s^2
Boys	55	82	29
Girls	60	86	46

139. In a survey of 80 males, 55 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. Conduct a hypothesis test.

140. Which of the following is preferable when designing a hypothesis test?

a. Maximize α and minimize β
b. Minimize α and maximize β
c. Maximize α and β
d. Minimize α and β

Use the following information to answer the next three exercises: 120 people were surveyed as to their favorite beverage (non-alcoholic). The results are below.

Beverage/Age	**0–9**	**10–19**	**20–29**	**30+**	**Totals**
Milk	14	10	6	0	30
Soda	3	8	26	15	52

Beverage/Age	**0–9**	**10–19**	**20–29**	**30+**	**Totals**
Juice	7	12	12	7	38
Totals	24	330	44	22	120

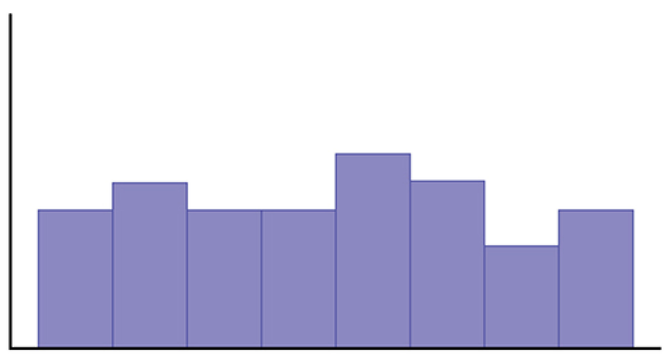
141. Are the events of milk and 30+:

a. independent events? Justify your answer.
b. mutually exclusive events? Justify your answer.

142. Suppose that one person is randomly chosen. Find the probability that person is 10–19 given that he or she prefers juice.

143. Are "Preferred Beverage" and "Age" independent events? Conduct a hypothesis test.

144. Given the following histogram, which distribution is the data most likely to come from?



a. uniform
b. exponential
c. normal
d. chi-square

Solutions

Chapter 3

1. c. parameter

2. a. population

3. b. statistic

4. d. sample

5. e. variable

6. quantitative continuous

7.

a. 2.27
b. 3.04
c. –1, 4, 4

8. Answers will vary.

Chapter 4

9. c. (0.80)(0.30)

10. b. No, and they are not mutually exclusive either.

11. a. all employed adult women

12. 0.5773

13. 0.0522

14. b. The middle fifty percent of the members lost from 2 to 8.5 lbs.

15. c. All of the data have the same value.

16. c. The lowest data value is the median.

17. 0.279

18. b. No, I expect to come out behind in money.

19. X = the number of patients calling in claiming to have the flu, who actually have the flu.
X = 0, 1, 2, …25

20. $B(25, 0.04)$

21. 0.0165

22. 1

23. c. quantitative discrete

24. all words used by Tom Clancy in his novels

Chapter 5

25.

a. 24%
b. 27%

26. qualitative

27. 0.36

28. 0.7636

29.

a. No
b. No

30. $B(10, 0.76)$

31. 0.9330

32.

a. X = the number of questions posted to the statistics listserv per day.
b. X = 0, 1, 2,…
c. $X \sim P(2)$
d. 0

33. $150

34. Matt

35.

a. false
b. true
c. false
d. false

36. 16

37. first quartile: 2
second quartile: 2
third quartile: 3

38. 0.5

39. $\frac{7}{15}$

40. $\frac{2}{15}$

Chapter 6

41.

a. true
b. true

c. False – the median and the mean are the same for this symmetric distribution.
d. true

42.

a. 8
b. 8
c. $P(x < k) = 0.65 = (k – 3)\left(\frac{1}{10}\right)$. $k = 9.5$

43.

a. False – $\frac{3}{4}$ of the data are at most five.
b. True – each quartile has 25% of the data.
c. False – that is unknown.
d. False – 50% of the data are four or less.

44. d. *G* and *H* are independent events.

45.

a. False – *J* and *K* are independent so they are not mutually exclusive which would imply dependency (meaning $P(J \text{ AND } K)$ is not 0).
b. False – see answer c.
c. True – $P(J \text{ OR } K) = P(J) + P(K) – P(J \text{ AND } K) = P(J) + P(K) – P(J)P(K) = 0.3 + 0.6 – (0.3)(0.6) = 0.72$. Note the $P(J \text{ AND } K) = P(J)P(K)$ because *J* and *K* are independent.
d. False – *J* and *K* are independent so $P(J) = P(J|K)$

46. a. $P(5)$

Chapter 7

47. a. $U(0, 4)$

48. b. 2 hour

49. a. $\frac{1}{4}$

50.

a. 0.7165
b. 4.16
c. 0

51. c. 5 years

52. c. exponential

53. 0.63

54. $B(14, 0.20)$

55. $B(14, 0.20)$

Chapter 8

56. c. the mean amount of weight lost by 15 people on the special weight loss diet.

57. 0.9951

58. 12.99

59. c. $\frac{1}{2}$

60. b. 0.60

61. c. $N(60, 5.477)$

62. 0.9990

63. a. eight days

64. c. 0.7500

65. a. 80%

66. b. 35%

67. b. no

68. b. quantitative continuous

69. c. 150

70. d. 0.06

71. c. 0.44

72. b. 0

Chapter 9

73. d. Matt is shorter than the average 14 year old boy.

74. Answers will vary.

75.

x	**Relative Frequency**	**Cumulative Relative Frequency**
1	0.3	0.3
2	0.2	0.2
4	0.4	0.4
5	0.1	0.1

76.

a. 2.8
b. 1.48
c. 90%

77. $M = 3$; $Q_1 = 1$; $Q_3 = 4$

78. 1 and 4

79. d. $\frac{8}{70}$

80. c. $\frac{40}{70}$

81. a. $\frac{9}{19}$

82. b. false

83. b. false

84. b. false

85.

a. X = the number of pies Lee bakes every day.
b. $P(20)$
c. 0.1122

86. CI: (5.25, 8.48)

87.

a. uniform
b. exponential
c. normal

Chapter 10

88. $\frac{77}{100}$

89. $\frac{12}{42}$

90.

a. false
b. false
c. true
d. false

91. $N(180, 16.43)$

92. a. The distribution for $\bar{X}$ is still uniform with the same mean and standard deviation as the distribution for *X*.

93. c. The distribution for $\sum X$ is normal with a larger mean and a larger standard deviation than the distribution for *X*.

94. $N\left(2, \frac{0.25}{\sqrt{16}}\right)$

95. Answers will vary.

96. 0.5000

97. 7.6

98. 5

99. 0.9431

Chapter 11

100. 7.5

101. 0.0122

102. $N(7, 0.63)$

103. 0.9911

104. b. Exponential

105.

a. true
b. false
c. false

106. Answers will vary.

107. Student's *t* with *df* = 15

108. (560.07, 719.93)

109. quantitative continuous data

110. quantitative discrete data

111.

a. *X* = the number of patients with a shotgun wound the emergency room gets per 28 days
b. *P*(4)
c. 0.0183

112. greater than

113. No; $P(x = 8) = 0.0348$

114. You will lose $5.

115. Becca

116. 14

117. Sample mean = 3.2
Sample standard deviation = 1.85
Median = 3
$Q_1 = 2$
$Q_3 = 5$
$IQR = 3$

118. d. $z = –1.19$
e. 0.1171
f. Do not reject the null hypothesis.

119. We conclude that the patient does have the HIV virus when, in fact, the patient does not.

120. c. $z = 2.21$; $p = 0.0136$
d. Reject the null hypothesis.
e. We conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is not greater.
f. We cannot conclude that the proportion of Californian professionals that wear jeans to work is greater than the proportion of non-Californian professionals when, in fact, it is greater.

121. c. dependent means

122. t_5

Chapter 12

123. (0.0424, 0.0770)

124. 2,401

125. Check student's solution.

126. 0.6321

127. $360

128. $N\left(72, \frac{72}{\sqrt{5}}\right)$

Chapter 13

129. 0.02

130. 0.40

131. $\frac{100}{140}$

132. $\frac{10}{60}$

133. *p*-value = 0; Reject the null hypothesis; conclude that they are dependent events

134. 8.4

135. *B*(14, 0.60)

136. d. Binomial

137. 0.3669

138. *p*-value = 0.0006; reject the null hypothesis; conclude that the averages are not equal

139. *p*-value = 0; reject the null hypothesis; conclude that the proportion of males is higher

140. Minimize α and β

141.

a. No
b. Yes, $P(M \text{ AND } 30+) = 0$

142. $\frac{12}{38}$

143. No; *p*-value = 0

144. a. uniform

References

Data from the *San Jose Mercury News*.

Baran, Daya. “20 Percent of Americans Have Never Used Email.” Webguild.org, 2010. Available online at: http://www.webguild.org/20080519/20-percent-of-americans-have-never-used-email (accessed October 17, 2013).

Data from *Parade Magazine*.

Practice Tests (1-4) and Final Exams

Practice Test 1

1.1: Definitions of Statistics, Probability, and Key Terms

Use the following information to answer the next three exercises. A grocery store is interested in how much money, on average, their customers spend each visit in the produce department. Using their store records, they draw a sample of 1,000 visits and calculate each customer's average spending on produce.

1. Identify the population, sample, parameter, statistic, variable, and data for this example.

a. population
b. sample
c. parameter
d. statistic
e. variable
f. data

2. What kind of data is "amount of money spent on produce per visit"?

a. qualitative
b. quantitative-continuous
c. quantitative-discrete

3. The study finds that the mean amount spent on produce per visit by the customers in the sample is $12.84. This is an example of a:

a. population
b. sample
c. parameter
d. statistic
e. variable

1.2: Data, Sampling, and Variation in Data and Sampling

Use the following information to answer the next two exercises. A health club is interested in knowing how many times a typical member uses the club in a week. They decide to ask every tenth customer on a specified day to complete a short survey including information about how many times they have visited the club in the past week.

4. What kind of a sampling design is this?

a. cluster
b. stratified
c. simple random
d. systematic

5. “Number of visits per week” is what kind of data?

a. qualitative
b. quantitative-continuous
c. quantitative-discrete

6. Describe a situation in which you would calculate a parameter, rather than a statistic.

7. The U.S. federal government conducts a survey of high school seniors concerning their plans for future education and employment. One question asks whether they are planning to attend a four-year college or university in the following year. Fifty percent answer yes to this question; that fifty percent is a:

a. parameter
b. statistic
c. variable
d. data

8. Imagine that the U.S. federal government had the means to survey all high school seniors in the U.S. concerning their plans for future education and employment, and found that 50 percent were planning to attend a 4-year college or university in the following year. This 50 percent is an example of a:

a. parameter
b. statistic
c. variable
d. data

Use the following information to answer the next three exercises. A survey of a random sample of 100 nurses working at a large hospital asked how many years they had been working in the profession. Their answers are summarized in the following (incomplete) table.

9. Fill in the blanks in the table and round your answers to two decimal places for the Relative Frequency and Cumulative Relative Frequency cells.

# of years	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
< 5	25		
5–10	30		
> 10	empty		

10. What proportion of nurses have five or more years of experience?

11. What proportion of nurses have ten or fewer years of experience?

12. Describe how you might draw a random sample of 30 students from a lecture class of 200 students.

13. Describe how you might draw a stratified sample of students from a college, where the strata are the students' class standing (freshman, sophomore, junior, or senior).

14. A manager wants to draw a sample, without replacement, of 30 employees from a workforce of 150. Describe how the chance of being selected will change over the course of drawing the sample.

15. The manager of a department store decides to measure employee satisfaction by selecting four departments at random, and conducting interviews with all the employees in those four departments. What type of survey design is this?

a. cluster
b. stratified
c. simple random
d. systematic

16. A popular American television sports program conducts a poll of viewers to see which team they believe will win the NFL (National Football League) championship this year. Viewers vote by calling a number displayed on the television screen and telling the operator which team they think will win. Do you think that those who participate in this poll are representative of all football fans in America?

17. Two researchers studying vaccination rates independently draw samples of 50 children, ages 3–18 months, from a large urban area, and determine if they are up to date on their vaccinations. One researcher finds that 84 percent of the children in her sample are up to date, and the other finds that 86 percent in his sample are up to date. Assuming both followed proper sampling procedures and did their calculations correctly, what is a likely explanation for this discrepancy?

18. A high school increased the length of the school day from 6.5 to 7.5 hours. Students who wished to attend this high school were required to sign contracts pledging to put forth their best effort on their school work and to obey the school rules; if they did not wish to do so, they could attend another high school in the district. At the end of one year, student performance on statewide tests had increased by ten percentage points over the previous year. Does this improvement prove that a longer school day improves student achievement?

19. You read a newspaper article reporting that eating almonds leads to increased life satisfaction. The study was conducted by the Almond Growers Association, and was based on a randomized survey asking people about their consumption of various foods, including almonds, and also about their satisfaction with different aspects of their life. Does anything about this poll lead you to question its conclusion?

20. Why is non-response a problem in surveys?

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. Compute the mean of the following numbers, and report your answer using one more decimal place than is present in the original data:
14, 5, 18, 23, 6

1.4: Experimental Design and Ethics

22. A psychologist is interested in whether the size of tableware (bowls, plates, etc.) influences how much college students eat. He randomly assigns 100 college students to one of two groups: the first is served a meal using normal-sized tableware, while the second is served the same meal, but using tableware that it 20 percent smaller than normal. He records how much food is consumed by each group. Identify the following components of this study.

a. population
b. sample
c. experimental units
d. explanatory variable
e. treatment
f. response variable

23. A researcher analyzes the results of the SAT (Scholastic Aptitude Test) over a five-year period and finds that male students on average score higher on the math section, and female students on average score higher on the verbal section. She concludes that these observed differences in test performance are due to genetic factors. Explain how lurking variables could offer an alternative explanation for the observed differences in test scores.

24. Explain why it would not be possible to use random assignment to study the health effects of smoking.

25. A professor conducts a telephone survey of a city's population by drawing a sample of numbers from the phone book and having her student assistants call each of the selected numbers once to administer the survey. What are some sources of bias with this survey?

26. A professor offers extra credit to students who take part in her research studies. What is an ethical problem with this method of recruiting subjects?

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

Use the following information to answer the next four exercises. The midterm grades on a chemistry exam, graded on a scale of 0 to 100, were:
62, 64, 65, 65, 68, 70, 72, 72, 74, 75, 75, 75, 76,78, 78, 81, 83, 83, 84, 85, 87, 88, 92, 95, 98, 98, 100, 100, 740

27. Do you see any outliers in this data? If so, how would you address the situation?

28. Construct a stem plot for this data, using only the values in the range 0–100.

29. Describe the distribution of exam scores.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. In a class of 35 students, seven students received scores in the 70–79 range. What is the relative frequency of scores in this range?

Use the following information to answer the next three exercises. You conduct a poll of 30 students to see how many classes they are taking this term. Your results are:
1; 1; 1; 1
2; 2; 2; 2; 2
3; 3; 3; 3; 3; 3; 3; 3
4; 4; 4; 4; 4; 4; 4; 4; 4
5; 5; 5; 5

31. You decide to construct a histogram of this data. What will be the range of your first bar, and what will be the central point?

32. What will be the widths and central points of the other bars?

33. Which bar in this histogram will be the tallest, and what will be its height?

34. You get data from the U.S. Census Bureau on the median household income for your city, and decide to display it graphically. Which is the better choice for this data, a bar graph or a histogram?

35. You collect data on the color of cars driven by students in your statistics class, and want to display this information graphically. Which is the better choice for this data, a bar graph or a histogram?

2.3: Measures of the Location of the Data

36. Your daughter brings home test scores showing that she scored in the 80th percentile in math and the 76th percentile in reading for her grade. Interpret these scores.

37. You have to wait 90 minutes in the emergency room of a hospital before you can see a doctor. You learn that your wait time was in the 82nd percentile of all wait times. Explain what this means, and whether you think it is good or bad.

2.4: Box Plots

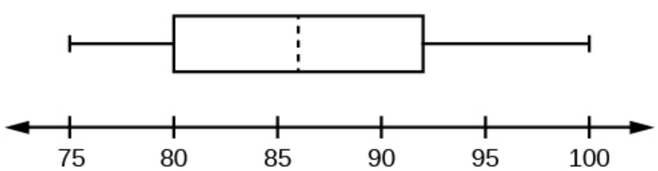
Use the following information to answer the next three exercises. 1; 1; 2; 3; 4; 4; 5; 5; 6; 7; 7; 8; 9

38. What is the median for this data?

39. What is the first quartile for this data?

40. What is the third quartile for this data?

Use the following information to answer the next four exercises. This box plot represents scores on the final exam for a physics class.



41. What is the median for this data, and how do you know?

42. What are the first and third quartiles for this data, and how do you know?

43. What is the interquartile range for this data?

44. What is the range for this data?

2.5: Measures of the Center of the Data

45. In a marathon, the median finishing time was 3:35:04 (three hours, 35 minutes, and four seconds). You finished in 3:34:10. Interpret the meaning of the median time, and discuss your time in relation to it.

Use the following information to answer the next three exercises. The value, in thousands of dollars, for houses on a block, are: 45; 47; 47.5; 51; 53.5; 125.

46. Calculate the mean for this data.

47. Calculate the median for this data.

48. Which do you think better reflects the average value of the homes on this block?

2.6: Skewness and the Mean, Median, and Mode

49. In a left-skewed distribution, which is greater?

a. the mean
b. the media
c. the mode

50. In a right-skewed distribution, which is greater?

a. the mean
b. the median
c. the mode

51. In a symmetrical distribution what will be the relationship among the mean, median, and mode?

2.7: Measures of the Spread of the Data

Use the following information to answer the next four exercises. 10; 11; 15; 15; 17; 22

52. Compute the mean and standard deviation for this data; use the sample formula for the standard deviation.

53. What number is two standard deviations above the mean of this data?

54. Express the number 13.7 in terms of the mean and standard deviation of this data.

55. In a biology class, the scores on the final exam were normally distributed, with a mean of 85, and a standard deviation of five. Susan got a final exam score of 95. Express her exam result as a z-score, and interpret its meaning.

3.1: Terminology

Use the following information to answer the next two exercises. You have a jar full of marbles: 50 are red, 25 are blue, and 15 are yellow. Assume you draw one marble at random for each trial, and replace it before the next trial.
Let $P(R)$ = the probability of drawing a red marble.
Let $P(B)$ = the probability of drawing a blue marble.
Let $P(Y)$ = the probability of drawing a yellow marble.

56. Find $P(B)$.

57. Which is more likely, drawing a red marble or a yellow marble? Justify your answer numerically.

Use the following information to answer the next two exercises. The following are probabilities describing a group of college students.
Let $P(M)$ = the probability that the student is male
Let $P(F)$ = the probability that the student is female
Let $P(E)$ = the probability the student is majoring in education
Let $P(S)$ = the probability the student is majoring in science

58. Write the symbols for the probability that a student, selected at random, is both female and a science major.

59. Write the symbols for the probability that the student is an education major, given that the student is male.

3.2: Independent and Mutually Exclusive Events

60. Events A and B are independent.
If $P(A) = 0.3$ and $P(B) = 0.5$, find $P(A \text{ AND } B)$.

61. C and D are mutually exclusive events.
If $P(C) = 0.18$ and $P(D) = 0.03$, find $P(C \text{ OR } D)$.

3.3: Two Basic Rules of Probability

62. In a high school graduating class of 300, 200 students are going to college, 40 are planning to work full-time, and 80 are taking a gap year. Are these events mutually exclusive?

Use the following information to answer the next two exercises. An archer hits the center of the target (the bullseye) 70 percent of the time. However, she is a streak shooter, and if she hits the center on one shot, her probability of hitting it on the shot immediately following is 0.85. Written in probability notation:
$P(A) = P(B) = P(\text{hitting the center on one shot}) = 0.70$
$P(B|A) = \text{P}(\text{hitting the center on a second shot, given that she hit it on the first}) = 0.85$

63. Calculate the probability that she will hit the center of the target on two consecutive shots.

64. Are $P(A)$ and $P(B)$ independent in this example?

3.4: Contingency Tables

Use the following information to answer the next three exercises. The following contingency table displays the number of students who report studying at least 15 hours per week, and how many made the honor roll in the past semester.

	Honor roll	**No honor roll**	**Total**
Study at least 15 hours/week		200	
Study less than 15 hours/week	125	193	
Total			1,000

65. Complete the table.

66. Find $P(\text{honor roll}|\text{study at least 15 hours per week})$.

67. What is the probability a student studies less than 15 hours per week?

68. Are the events “study at least 15 hours per week” and “makes the honor roll” independent? Justify your answer numerically.

3.5: Tree and Venn Diagrams

69. At a high school, some students play on the tennis team, some play on the soccer team, but neither plays both tennis and soccer. Draw a Venn diagram illustrating this.

70. At a high school, some students play tennis, some play soccer, and some play both. Draw a Venn diagram illustrating this.

Practice Test 1 Solutions

1.1: Definitions of Statistics, Probability, and Key Terms

1.

a. population: all the shopping visits by all the store’s customers
b. sample: the 1,000 visits drawn for the study
c. parameter: the average expenditure on produce per visit by all the store’s customers
d. statistic: the average expenditure on produce per visit by the sample of 1,000
e. variable: the expenditure on produce for each visit
f. data: the dollar amounts spent on produce; for instance, $15.40, $11.53, etc

2. c

3. d

1.2: Data, Sampling, and Variation in Data and Sampling

4. d

5. c

6. Answers will vary.
Sample Answer: Any solution in which you use data from the entire population is acceptable. For instance, a professor might calculate the average exam score for her class: because the scores of all members of the class were used in the calculation, the average is a parameter.

7. b

8. a

9.

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
< 5	25	0.25	0.25

# of years	Frequency	Relative Frequency	Cumulative Relative Frequency
5–10	30	0.30	0.55
> 10	45	0.45	1.00

10. 0.75

11. 0.55

12. Answers will vary.
Sample Answer: One possibility is to obtain the class roster and assign each student a number from 1 to 200. Then use a random number generator or table of random number to generate 30 numbers between 1 and 200, and select the students matching the random numbers. It would also be acceptable to write each student's name on a card, shuffle them in a box, and draw 30 names at random.

13. One possibility would be to obtain a roster of students enrolled in the college, including the class standing for each student. Then you would draw a proportionate random sample from within each class (for instance, if 30 percent of the students in the college are freshman, then 30 percent of your sample would be drawn from the freshman class).

14. For the first person picked, the chance of any individual being selected is one in 150. For the second person, it is one in 149, for the third it is one in 148, and so on. For the 30th person selected, the chance of selection is one in 121.

15. a

16. No. There are at least two chances for bias. First, the viewers of this particular program may not be representative of American football fans as a whole. Second, the sample will be self-selected, because people have to make a phone call in order to take part, and those people are probably not representative of the American football fan population as a whole.

17. These results (84 percent in one sample, 86 percent in the other) are probably due to sampling variability. Each researcher drew a different sample of children, and you would not expect them to get exactly the same result, although you would expect the results to be similar, as they are in this case.

18. No. The improvement could also be due to self-selection: only motivated students were willing to sign the contract, and they would have done well even in a school with 6.5 hour days. Because both changes were implemented at the same time, it is not possible to separate out their influence.

19. At least two aspects of this poll are troublesome. The first is that it was conducted by a group who would benefit by the result—almond sales are likely to increase if people believe that eating almonds will make them happier. The second is that this poll found that almond consumption and life satisfaction are correlated, but does not establish that eating almonds causes satisfaction. It is equally possible, for instance, that people with higher incomes are more likely to eat almonds, and are also more satisfied with their lives.

20. You want the sample of people who take part in a survey to be representative of the population from which they are drawn. People who refuse to take part in a survey often have different views than those who do participate, and so even a random sample may produce biased results if a large percentage of those selected refuse to participate in a survey.

1.3: Frequency, Frequency Tables, and Levels of Measurement

21. 13.2

1.4: Experimental Design and Ethics

22.

a. population: all college students
b. sample: the 100 college students in the study
c. experimental units: each individual college student who participated
d. explanatory variable: the size of the tableware
e. treatment: tableware that is 20 percent smaller than normal
f. response variable: the amount of food eaten

23. There are many lurking variables that could influence the observed differences in test scores. Perhaps the boys, on average, have taken more math courses than the girls, and the girls have taken more English classes than the boys. Perhaps the boys have been encouraged by their families and teachers to prepare for a career in math and science, and thus have put more effort into studying math, while the girls have been encouraged to prepare for fields like communication and psychology that are more focused on language use. A study design would have to control for these and other potential lurking variables (anything that could explain the observed difference in test scores, other than the genetic explanation) in order to draw a scientifically sound conclusion about genetic differences.

24. To use random assignment, you would have to be able to assign people to either smoke or not smoke. Because smoking has many harmful effects, this would not be an ethical experiment. Instead, we study people who have chosen to smoke, and compare them to others who have chosen not to smoke, and try to control for the other ways those two groups may differ (lurking variables).

25. Sources of bias include the fact that not everyone has a telephone, that cell phone numbers are often not listed in published directories, and that an individual might not be at home at the time of the phone call; all these factors make it likely that the respondents to the survey will not be representative of the population as a whole.

26. Research subjects should not be coerced into participation, and offering extra credit in exchange for participation could be construed as coercion. In addition, this method will result in a volunteer sample, which cannot be assumed to be representative of the population as a whole.

2.1: Stem-and Leaf Graphs (Stemplots), Line Graphs, and Bar Graphs

27. The value 740 is an outlier, because the exams were graded on a scale of 0 to 100, and 740 is far outside that range. It may be a data entry error, with the actual score being 74, so the professor should check that exam again to see what the actual score was.

28.

Stem	Leaf
6	2 4 5 5 8
7	0 2 2 4 5 5 5 6 8 8
8	1 3 3 4 5 7 8
9	2 5 8 8

Stem	Leaf
10	0 0

29. Most scores on this exam were in the range of 70–89, with a few scoring in the 60–69 range, and a few in the 90–100 range.

2.2: Histograms, Frequency Polygons, and Time Series Graphs

30. $RF = \frac{7}{35} = 0.2$

31. The range will be 0.5–1.5, and the central point will be 1.

32. Range 1.5–2.5, central point 2; range 2.5–3.5, central point 3; range 3.5–4.5, central point 4; range 4.5–5.5., central point 5.

33. The bar from 3.5 to 4.5, with a central point of 4, will be tallest; its height will be nine, because there are nine students taking four courses.

34. The histogram is a better choice, because income is a continuous variable.

35. A bar graph is the better choice, because this data is categorical rather than continuous.

2.3: Measures of the Location of the Data

36. Your daughter scored better than 80 percent of the students in her grade on math and better than 76 percent of the students in reading. Both scores are very good, and place her in the upper quartile, but her math score is slightly better in relation to her peers than her reading score.

37. You had an unusually long wait time, which is bad: 82 percent of patients had a shorter wait time than you, and only 18 percent had a longer wait time.

2.4: Box Plots

38. 5

39. 3

40. 7

41. The median is 86, as represented by the vertical line in the box.

42. The first quartile is 80, and the third quartile is 92, as represented by the left and right boundaries of the box.

43. IQR = 92 – 80 = 12

44. Range = 100 – 75 = 25

2.5: Measures of the Center of the Data

45. Half the runners who finished the marathon ran a time faster than 3:35:04, and half ran a time slower than 3:35:04. Your time is faster than the median time, so you did better than more than half of the runners in this race.

46. 61.5, or $61,500

47. 49.25 or $49,250

48. The median, because the mean is distorted by the high value of one house.

2.6: Skewness and the Mean, Median, and Mode

49. c

50. a

51. They will all be fairly close to each other.

2.7: Measures of the Spread of the Data

52. Mean: 15
Standard deviation: 4.3

$\mu = \frac{10+11+15+15+17+22}{6} = 15$

$s = \sqrt{\frac{\sum(x-x)^2}{n-1}} = \sqrt{\frac{94}{5}} = 4.3$

53. 15 + (2)(4.3) = 23.6

54. 13.7 is one standard deviation below the mean of this data, because 15 – 4.3 = 10.7

55. $z = \frac{95-85}{5} = 2.0$
Susan's *z*-score was 2.0, meaning she scored two standard deviations above the class mean for the final exam.

3.1: Terminology

56. $P(B) = \frac{25}{90} = 0.28$

57. Drawing a red marble is more likely.
$P(R) = \frac{50}{80} = 0.62$
$P(Y) = \frac{15}{80} = 0.19$

58. $P(F \text{ AND } S)$

59. $P(E|M)$

3.2: Independent and Mutually Exclusive Events

60. $P(A \text{ AND } B)$ = (0.3)(0.5) = 0.15

61. $P(C \text{ OR } D) = 0.18 + 0.03 = 0.21$

3.3: Two Basic Rules of Probability

62. No, they cannot be mutually exclusive, because they add up to more than 300. Therefore, some students must fit into two or more categories (e.g., both going to college and working full time).

63. $P(A \text{ and } B) = (P(B|A))(P(A)) = (0.85)(0.70) = 0.595$

64. No. If they were independent, $P(B)$ would be the same as $P(B|A)$. We know this is not the case, because $P(B) = 0.70$ and $P(B|A) = 0.85$.

3.4: Contingency Tables

65.

	Honor roll	**No honor roll**	**Total**
Study at least 15 hours/week	482	200	682
Study less than 15 hours/week	125	193	318
Total	607	393	1,000

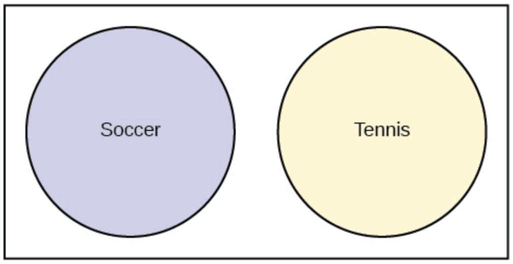
66. $P(\text{honor roll}|\text{study at least 15 hours word per week}) = \frac{482}{1000} = 0.482$

67. $P(\text{studies less than 15 hours word per week}) = \frac{125+193}{1000} = 0.318$

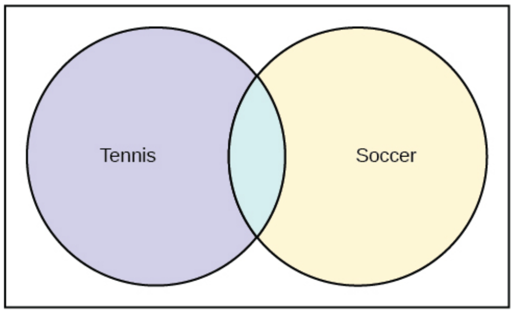
68. Let $P(S)$ = study at least 15 hours per week
Let $P(H)$ = makes the honor roll
From the table, $P(S) = 0.682$, $P(H) = 0.607$, and $P(S \text{ AND } H) = 0.482$.
If $P(S)$ and $P(H)$ were independent, then $P(S \text{ AND } H)$ would equal $(P(S))(P(H))$.
However, $(P(S))(P(H)) = (0.682)(0.607) = 0.414$, while $P(S \text{ AND } H) = 0.482$.
Therefore, $P(S)$ and $P(H)$ are not independent.

3.5: Tree and Venn Diagrams

69.



70.



Practice Test 2

4.1: Probability Distribution Function (PDF) for a Discrete Random Variable

Use the following information to answer the next five exercises. You conduct a survey among a random sample of students at a particular university. The data collected includes their major, the number of classes they took the previous semester, and amount of money they spent on books purchased for classes in the previous semester.

1. If X = student's major, then what is the domain of X?

2. If Y = the number of classes taken in the previous semester, what is the domain of Y?

3. If Z = the amount of money spent on books in the previous semester, what is the domain of Z?

4. Why are X, Y, and Z in the previous example random variables?

5. After collecting data, you find that for one case, $z = –7$. Is this a possible value for Z?

6. What are the two essential characteristics of a discrete probability distribution?

Use this discrete probability distribution represented in this table to answer the following six questions. The university library records the number of books checked out by each patron over the course of one day, with the following result:

x	***P*(*x*)**

x	*P*(*x*)
0	0.20
1	0.45
2	0.20
3	0.10
4	0.05

7. Define the random variable X for this example.

8. What is $P(x > 2)$?

9. What is the probability that a patron will check out at least one book?

10. What is the probability a patron will take out no more than three books?

11. If the table listed $P(x)$ as 0.15, how would you know that there was a mistake?

12. What is the average number of books taken out by a patron?

4.2: Mean or Expected Value and Standard Deviation

Use the following information to answer the next four exercises. Three jobs are open in a company: one in the accounting department, one in the human resources department, and one in the sales department. The accounting job receives 30 applicants, and the human resources and sales department 60 applicants.

13. If X = the number of applications for a job, use this information to fill in [link].

x	*P*(*x*)	*xP*(*x*)

14. What is the mean number of applicants?

15. What is the PDF for X?

16. Add a fourth column to the table, for $(x - \mu)^2P(x)$.

17. What is the standard deviation of X?

4.3: Binomial Distribution

18. In a binomial experiment, if $p = 0.65$, what does q equal?

19. What are the required characteristics of a binomial experiment?

20. Joe conducts an experiment to see how many times he has to flip a coin before he gets four heads in a row. Does this qualify as a binomial experiment?

Use the following information to answer the next three exercises. In a particularly community, 65 percent of households include at least one person who has graduated from college. You randomly sample 100 households in this community. Let X = the number of households including at least one college graduate.

21. Describe the probability distribution of X.

22. What is the mean of X?

23. What is the standard deviation of X?

Use the following information to answer the next four exercises. Joe is the star of his school's baseball team. His batting average is 0.400, meaning that for every ten times he comes to bat (an at-bat), four of those times he gets a hit. You decide to track his batting performance his next 20 at-bats.

24. Define the random variable X in this experiment.

25. Assuming Joe's probability of getting a hit is independent and identical across all 20 at-bats, describe the distribution of X.

26. Given this information, what number of hits do you predict Joe will get?

27. What is the standard deviation of X?

4.4: Geometric Distribution

28. What are the three major characteristics of a geometric experiment?

29. You decide to conduct a geometric experiment by flipping a coin until it comes up heads. This takes five trials. Represent the outcomes of this trial, using H for heads and T for tails.

30. You are conducting a geometric experiment by drawing cards from a normal 52-card pack, with replacement, until you draw the Queen of Hearts. What is the domain of X for this experiment?

31. You are conducting a geometric experiment by drawing cards from a normal 52-card deck, without replacement, until you draw a red card. What is the domain of X for this experiment?

Use the following information to answer the next three exercises. In a particular university, 27 percent of students are engineering majors. You decide to select students at random until you choose one that is an engineering major. Let X = the number of students you select until you find one that is an engineering major.

32. What is the probability distribution of X?

33. What is the mean of X?

34. What is the standard deviation of X?

4.5: Hypergeometric Distribution

35. You draw a random sample of ten students to participate in a survey, from a group of 30, consisting of 16 boys and 14 girls. You are interested in the probability that seven of the students chosen will be boys. Does this qualify as a hypergeometric experiment? List the conditions and whether or not they are met.

36. You draw five cards, without replacement, from a normal 52-card deck of playing cards, and are interested in the probability that two of the cards are spades. What are the group of interest, size of the group of interest, and sample size for this example?

4.6: Poisson Distribution

37. What are the key characteristics of the Poisson distribution?

Use the following information to answer the next three exercises. The number of drivers to arrive at a toll booth in an hour can be modeled by the Poisson distribution.

38. If X = the number of drivers, and the average numbers of drivers per hour is four, how would you express this distribution?

39. What is the domain of X?

40. What are the mean and standard deviation of X?

5.1: Continuous Probability Functions

41. You conduct a survey of students to see how many books they purchased the previous semester, the total amount they paid for those books, the number they sold after the semester was over, and the amount of money they received for the books they sold. Which variables in this survey are discrete, and which are continuous?

42. With continuous random variables, we never calculate the probability that X has a particular value, but always speak in terms of the probability that X has a value within a particular range. Why is this?

43. For a continuous random variable, why are $P(x < c)$ and $P(x \leq c)$ equivalent statements?

44. For a continuous probability function, $P(x < 5) = 0.35$. What is $P(x > 5)$, and how do you know?

45. Describe how you would draw the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$. What type of a distribution is this?

46. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(0 < x < 4)$?

5.2: The Uniform Distribution

47. For the continuous probability distribution described by the function $f(x) = \frac{1}{10}$ for $0 \leq x \leq 10$, what is the $P(2 < x < 5)$?

Use the following information to answer the next four exercises. The number of minutes that a patient waits at a medical clinic to see a doctor is represented by a uniform distribution between zero and 30 minutes, inclusive.

48. If X equals the number of minutes a person waits, what is the distribution of X?

49. Write the probability density function for this distribution.

50. What is the mean and standard deviation for waiting time?

51. What is the probability that a patient waits less than ten minutes?

5.3: The Exponential Distribution

52. The distribution of the variable *X*, representing the average time to failure for an automobile battery, can be written as: $X \sim Exp(m)$. Describe this distribution in words.

53. If the value of *m* for an exponential distribution is ten, what are the mean and standard deviation for the distribution?

54. Write the probability density function for a variable distributed as: $X \sim Exp(0.2)$.

6.1: The Standard Normal Distribution

55. Translate this statement about the distribution of a random variable *X* into words: $X \sim (100, 15)$.

56. If the variable *X* has the standard normal distribution, express this symbolically.

Use the following information for the next six exercises. According to the World Health Organization, distribution of height in centimeters for girls aged five years and no months has the distribution: $X \sim N(109, 4.5)$.

57. What is the *z*-score for a height of 112 inches?

58. What is the *z*-score for a height of 100 centimeters?

59. Find the *z*-score for a height of 105 centimeters and explain what that means In the context of the population.

60. What height corresponds to a *z*-score of 1.5 in this population?

61. Using the empirical rule, we expect about 68 percent of the values in a normal distribution to lie within one standard deviation above or below the mean. What does this mean, in terms of a specific range of values, for this distribution?

62. Using the empirical rule, about what percent of heights in this distribution do you expect to be between 95.5 cm and 122.5 cm?

6.2: Using the Normal Distribution

Use the following information to answer the next four exercises. The distributor of lotto tickets claims that 20 percent of the tickets are winners. You draw a sample of 500 tickets to test this proposition.

63. Can you use the normal approximation to the binomial for your calculations? Why or why not.

64. What are the expected mean and standard deviation for your sample, assuming the distributor’s claim is true?

65. What is the probability that your sample will have a mean greater than 100?

66. If the *z*-score for your sample result is –2.00, explain what this means, using the empirical rule.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. What does the central limit theorem state with regard to the distribution of sample means?

68. The distribution of results from flipping a fair coin is uniform: heads and tails are equally likely on any flip, and over a large number of trials, you expect about the same number of heads and tails. Yet if you conduct a study by flipping 30 coins and recording the number of heads, and repeat this 100 times, the distribution of the mean number of heads will be approximately normal. How is this possible?

69. The mean of a normally-distributed population is 50, and the standard deviation is four. If you draw 100 samples of size 40 from this population, describe what you would expect to see in terms of the sampling distribution of the sample mean.

70. X is a random variable with a mean of 25 and a standard deviation of two. Write the distribution for the sample mean of samples of size 100 drawn from this population.

71. Your friend is doing an experiment drawing samples of size 50 from a population with a mean of 117 and a standard deviation of 16. This sample size is large enough to allow use of the central limit theorem, so he says the standard deviation of the sampling distribution of sample means will also be 16. Explain why this is wrong, and calculate the correct value.

72. You are reading a research article that refers to “the standard error of the mean.” What does this mean, and how is it calculated?

Use the following information to answer the next six exercises. You repeatedly draw samples of $n = 100$ from a population with a mean of 75 and a standard deviation of 4.5.

73. What is the expected distribution of the sample means?

74. One of your friends tries to convince you that the standard error of the mean should be 4.5. Explain what error your friend made.

75. What is the z-score for a sample mean of 76?

76. What is the z-score for a sample mean of 74.7?

77. What sample mean corresponds to a z-score of 1.5?

78. If you decrease the sample size to 50, will the standard error of the mean be smaller or larger? What would be its value?

Use the following information to answer the next two questions. We use the empirical rule to analyze data for samples of size 60 drawn from a population with a mean of 70 and a standard deviation of 9.

79. What range of values would you expect to include 68 percent of the sample means?

80. If you increased the sample size to 100, what range would you expect to contain 68 percent of the sample means, applying the empirical rule?

7.2: The Central Limit Theorem for Sums

81. How does the central limit theorem apply to sums of random variables?

82. Explain how the rules applying the central limit theorem to sample means, and to sums of a random variable, are similar.

83. If you repeatedly draw samples of size 50 from a population with a mean of 80 and a standard deviation of four, and calculate the sum of each sample, what is the expected distribution of these sums?

Use the following information to answer the next four exercises. You draw one sample of size 40 from a population with a mean of 125 and a standard deviation of seven.

84. Compute the sum. What is the probability that the sum for your sample will be less than 5,000?

85. If you drew samples of this size repeatedly, computing the sum each time, what range of values would you expect to contain 95 percent of the sample sums?

86. What value is one standard deviation below the mean?

87. What value corresponds to a *z*-score of 2.2?

7.3: Using the Central Limit Theorem

88. What does the law of large numbers say about the relationship between the sample mean and the population mean?

89. Applying the law of large numbers, which sample mean would expect to be closer to the population mean, a sample of size ten or a sample of size 100?

Use this information for the next three questions. A manufacturer makes screws with a mean diameter of 0.15 cm (centimeters) and a range of 0.10 cm to 0.20 cm; within that range, the distribution is uniform.

90. If *X* = the diameter of one screw, what is the distribution of *X*?

91. Suppose you repeatedly draw samples of size 100 and calculate their mean. Applying the central limit theorem, what is the distribution of these sample means?

92. Suppose you repeatedly draw samples of 60 and calculate their sum. Applying the central limit theorem, what is the distribution of these sample sums?

Practice Test 2 Solutions

Probability Distribution Function (PDF) for a Discrete Random Variable

1. The domain of *X* = {English, Mathematics,….], i.e., a list of all the majors offered at the university, plus "undeclared."

2. The domain of *Y* = {0, 1, 2, …}, i.e., the integers from 0 to the upper limit of classes allowed by the university.

3. The domain of *Z* = any amount of money from 0 upwards.

4. Because they can take any value within their domain, and their value for any particular case is not known until the survey is completed.

5. No, because the domain of *Z* includes only positive numbers (you can't spend a negative amount of money). Possibly the value –7 is a data entry error, or a special code to indicated that the student did not answer the question.

6. The probabilities must sum to 1.0, and the probabilities of each event must be between 0 and 1, inclusive.

7. Let *X* = the number of books checked out by a patron.

8. $P(x > 2) = 0.10 + 0.05 = 0.15$

9. $P(x \geq 0) = 1 – 0.20 = 0.80$

10. $P(x \leq 3) = 1 – 0.05 = 0.95$

11. The probabilities would sum to 1.10, and the total probability in a distribution must always equal 1.0.

12. x = 0(0.20) + 1(0.45) + 2(0.20) + 3(0.10) + 4(0.05) = 1.35

Mean or Expected Value and Standard Deviation

13.

x	***P*(*x*)**	***xP*(*x*)**
30	0.33	9.90
40	0.33	13.20
60	0.33	19.80

14. x = 9.90 + 13.20 + 19.80 = 42.90

15. $P(x = 30) = 0.33$
$P(x = 40) = 0.33$
$P(x = 60) = 0.33$

16.

x	***P*(*x*)**	***xP*(*x*)**	**$(x – \mu)^2 P(x)$**
30	0.33	9.90	$(30 – 42.90)^2(0.33) = 54.91$
40	0.33	13.20	$(40 – 42.90)^2(0.33) = 2.78$
60	0.33	19.90	$(60 – 42.90)^2(0.33) = 96.49$

17. $\sigma_x = \sqrt{54.91 + 2.78 + 96.49} = 12.42$

Binomial Distribution

18. $q = 1 – 0.65 = 0.35$

19.

1. There are a fixed number of trials.
2. There are only two possible outcomes, and they add up to 1.
3. The trials are independent and conducted under identical conditions.

20. No, because there are not a fixed number of trials

21. *X* ~ *B*(100, 0.65)

22. μ = *np* = 100(0.65) = 65

23. $\sigma_x = \sqrt{npq} = \sqrt{100(0.65)(0.35)} = 4.77$

24. *X* = Joe gets a hit in one at-bat (in one occasion of his coming to bat)

25. *X* ~ *B*(20, 0.4)

26. μ = *np* = 20(0.4) = 8

27.$\sigma_x = \sqrt{npq} = \sqrt{20(0.40)(0.60)} = 2.19$

4.4: Geometric Distribution

28.

1. A series of Bernoulli trials are conducted until one is a success, and then the experiment stops.
2. At least one trial is conducted, but there is no upper limit to the number of trials.
3. The probability of success or failure is the same for each trial.

29. *T T T T H*

30. The domain of *X* = {1, 2, 3, 4, 5, ….n}. Because you are drawing with replacement, there is no upper bound to the number of draws that may be necessary.

31. The domain of *X* = {1, 2, 3, 4, 5, 6, 7, 8., 9, 10, 11, 12…27}. Because you are drawing without replacement, and 26 of the 52 cards are red, you have to draw a red card within the first 17 draws.

32. *X* ~ *G*(0.24)

33. $\mu = \frac{1}{p} = \frac{1}{0.27} = 3.70$

34. $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.27}{0.27^2}} = 3.16$

4.5: Hypergeometric Distribution

35. Yes, because you are sampling from a population composed of two groups (boys and girls), have a group of interest (boys), and are sampling without replacement (hence, the probabilities change with each pick, and you are not performing Bernoulli trials).

36. The group of interest is the cards that are spades, the size of the group of interest is 13, and the sample size is five.

4.6: Poisson Distribution

37. A Poisson distribution models the number of events occurring in a fixed interval of time or space, when the events are independent and the average rate of the events is known.

38. $X \sim P(4)$

39. The domain of X = {0, 1, 2, 3, …..) i.e., any integer from 0 upwards.

40. $\mu = 4$
$\sigma = \sqrt{4} = 2$

5.1: Continuous Probability Functions

41. The discrete variables are the number of books purchased, and the number of books sold after the end of the semester. The continuous variables are the amount of money spent for the books, and the amount of money received when they were sold.

42. Because for a continuous random variable, $P(x = c) = 0$, where c is any single value. Instead, we calculate $P(c < x < d)$, i.e., the probability that the value of x is between the values c and d.

43. Because $P(x = c) = 0$ for any continuous random variable.

44. $P(x > 5) = 1 – 0.35 = 0.65$, because the total probability of a continuous probability function is always 1.

45. This is a uniform probability distribution. You would draw it as a rectangle with the vertical sides at 0 and 20, and the horizontal sides at $\frac{1}{10}$ and 0.

46. $P\left(0\ < x < 4\right) = (4 - 0)\left(\frac{1}{10}\right) =\ 0.4$

5.2: The Uniform Distribution

47. $P\left(2\ < x < 5\right) = (5 - 2)\left(\frac{1}{10}\right) =\ 0.3$

48. $X \sim U(0, 15)$

49. $f(x) = \frac{1}{b-a}$ for $(a \le x \le b)$ so $f(x) = \frac{1}{30}$ for $(0 \le x \le 30)$

50. $\mu =\ \frac{a+b}{2}\ =\ \frac{0+30}{5}\ = 15.0$

$\sigma =\ \sqrt{\frac{(b-a)^2}{12}}\ =\ \sqrt{\frac{(30-0)^2}{12}} = 8.66$

51. $P\left(x < 10\right) = (10)\left(\frac{1}{30}\right) =\ 0.33$

5.3: The Exponential Distribution

52. X has an exponential distribution with decay parameter m and mean and standard deviation $\frac{1}{m}$. In this distribution, there will be a relatively large numbers of small values, with values becoming less common as they become larger.

53. $\mu = \sigma = \frac{1}{m} = \frac{1}{10} = 0.1$

54. $f(x) = 0.2e^{-0.2x}$ where $x \ge 0$.

6.1: The Standard Normal Distribution

55. The random variable *X* has a normal distribution with a mean of 100 and a standard deviation of 15.

56. *X* ~ *N*(0,1)

57. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{112-109}{4.5} = 0.67$

58. $z = \frac{x-\mu}{\sigma}$ so $z = \frac{100-109}{4.5} = -2.00$

59. $z = \frac{105-109}{4.5} = -0.89$
This girl is shorter than average for her age, by 0.89 standard deviations.

60. 109 + (1.5)(4.5) = 115.75 cm

61. We expect about 68 percent of the heights of girls of age five years and zero months to be between 104.5 cm and 113.5 cm.

62. We expect 99.7 percent of the heights in this distribution to be between 95.5 cm and 122.5 cm, because that range represents the values three standard deviations above and below the mean.

6.2: Using the Normal Distribution

63. Yes, because both np and nq are greater than five.
np = (500)(0.20) = 100 and *nq* = 500(0.80) = 400

64. $\mu = np = (500)(0.20) = 100$

$\sigma = \sqrt{npq} = \sqrt{500(0.20)(0.80)} = 8.94$

65. Fifty percent, because in a normal distribution, half the values lie above the mean.

66. The results of our sample were two standard deviations below the mean, suggesting it is unlikely that 20 percent of the lotto tickets are winners, as claimed by the distributor, and that the true percent of winners is lower. Applying the Empirical Rule, If that claim were true, we would expect to see a result this far below the mean only about 2.5 percent of the time.

7.1: The Central Limit Theorem for Sample Means (Averages)

67. The central limit theorem states that if samples of sufficient size drawn from a population, the distribution of sample means will be normal, even if the distribution of the population is not normal.

68. The sample size of 30 is sufficiently large in this example to apply the central limit theorem. This theorem] states that for samples of sufficient size drawn from a population, the sampling distribution of the sample mean will approach normality, regardless of the distribution of the population from which the samples were drawn.

69. You would not expect each sample to have a mean of 50, because of sampling variability. However, you would expect the sampling distribution of the sample means to cluster around 50, with an approximately normal distribution, so that values close to 50 are more common than values further removed from 50.

70. $X \sim N(25, 0.2)$ because $X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

71. The standard deviation of the sampling distribution of the sample means can be calculated using the formula $\left(\frac{\sigma_x}{\sqrt{n}}\right)$, which in this case is $\left(\frac{16}{\sqrt{50}}\right)$. The correct value for the standard deviation of the sampling distribution of the sample means is therefore 2.26.

72. The standard error of the mean is another name for the standard deviation of the sampling distribution of the sample mean. Given samples of size n drawn from a population with standard deviation σ_x, the standard error of the mean is $\left(\frac{\sigma_x}{\sqrt{n}}\right)$.

73. $X \sim N(75, 0.45)$

74. Your friend forgot to divide the standard deviation by the square root of *n*.

75. $z = \frac{x - \mu_x}{\sigma_x} = \frac{76-75}{4.5} = 2.2$

76. $z = \frac{x - \mu_x}{\sigma_x} = \frac{74.7-75}{4.5} = -0.67$

77. 75 + (1.5)(0.45) = 75.675

78. The standard error of the mean will be larger, because you will be dividing by a smaller number. The standard error of the mean for samples of size *n* = 50 is:
$\left(\frac{\sigma_x}{\sqrt{n}}\right) = \frac{4.5}{\sqrt{50}} = 0.64$

79. You would expect this range to include values up to one standard deviation above or below the mean of the sample means. In this case:
$70 + \frac{9}{\sqrt{60}} = 71.16$ and $70 - \frac{9}{\sqrt{60}} = 68.84$ so you would expect 68 percent of the sample means to be between 68.84 and 71.16.

80. $70 + \frac{9}{\sqrt{100}} = 70.9$ and $70 - \frac{9}{\sqrt{100}} = 69.1$ so you would expect 68 percent of the sample means to be between 69.1 and 70.9. Note that this is a narrower interval due to the increased sample size.

7.2: The Central Limit Theorem for Sums

81. For a random variable *X*, the random variable *ΣX* will tend to become normally distributed as the size n of the samples used to compute the sum increases.

82. Both rules state that the distribution of a quantity (the mean or the sum) calculated on samples drawn from a population will tend to have a normal distribution, as the sample size increases, regardless of the distribution of population from which the samples are drawn.

83. $\Sigma X \sim N\left(n\mu_x, \left(\sqrt{n}\right)(\sigma_x)\right)$ so $\Sigma X \sim N(4000, 28.3)$

84.The probability is 0.50, because 5,000 is the mean of the sampling distribution of sums of size 40 from this population. Sums of random variables computed from a sample of sufficient size are normally distributed, and in a normal distribution, half the values lie below the mean.

85. Using the empirical rule, you would expect 95 percent of the values to be within two standard deviations of the mean. Using the formula for the standard deviation is for a sample sum: $\left(\sqrt{n}\right)(\sigma_x) = \left(\sqrt{40}\right)(7) = 44.3$ so you would expect 95 percent of the values to be between 5,000 + (2)(44.3) and 5,000 – (2)(44.3), or between 4,911.4 and 588.6.

86. $\mu - \left(\sqrt{n}\right)(\sigma_x) = 5000 - \left(\sqrt{40}\right)(7) = 4955.7$

87. $5000 + (2.2)\left(\sqrt{40}\right)(7) = 5097.4$

7.3: Using the Central Limit Theorem

88. The law of large numbers says that as sample size increases, the sample mean tends to get nearer and nearer to the population mean.

89. You would expect the mean from a sample of size 100 to be nearer to the population mean, because the law of large numbers says that as sample size increases, the sample mean tends to approach the population mea.

90. $X \sim N(0.10, 0.20)$

91. $X \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$ and the standard deviation of a uniform distribution is $\frac{b-a}{\sqrt{12}}$. In this example, the standard deviation of the distribution is $\frac{b-a}{\sqrt{12}} = \frac{0.10}{\sqrt{12}} = 0.03$
so $X \sim N\left(0.15, 0.003\right)$

92. $\Sigma X \sim N((n)(\mu_x), (\sqrt{n})(\sigma_x))$ so $\Sigma X \sim N(9.0, 0.23)$

Practice Test 3

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

Use the following information to answer the next seven exercises. You draw a sample of size 30 from a normally distributed population with a standard deviation of four.

1. What is the standard error of the sample mean in this scenario, rounded to two decimal places?

2. What is the distribution of the sample mean?

3. If you want to construct a two-sided 95% confidence interval, how much probability will be in each tail of the distribution?

4. What is the appropriate *z*-score and error bound or margin of error (*EBM*) for a 95% confidence interval for this data?

5. Rounding to two decimal places, what is the 95% confidence interval if the sample mean is 41?

6. What is the 90% confidence interval if the sample mean is 41? Round to two decimal places

7. Suppose the sample size in this study had been 50, rather than 30. What would the 95% confidence interval be if the sample mean is 41? Round your answer to two decimal places.

8. For any given data set and sampling situation, which would you expect to be wider: a 95% confidence interval or a 99% confidence interval?

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's *t*

9. Comparing graphs of the standard normal distribution (*z*-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how do they differ?

10. Comparing graphs of the standard normal distribution (*z*-distribution) and a *t*-distribution with 15 degrees of freedom (*df*), how are they similar?

Use the following information to answer the next five exercises. Body temperature is known to be distributed normally among healthy adults. Because you do not know the population standard deviation, you use the t-distribution to study body temperature. You collect data from a random sample of 20 healthy adults and find that your sample temperatures have a mean of 98.4 and a sample standard deviation of 0.3 (both in degrees Fahrenheit).

11. What is the degrees of freedom (*df*) for this study?

12. For a two-tailed 95% confidence interval, what is the appropriate *t*-value to use in the formula?

13. What is the 95% confidence interval?

14. What is the 99% confidence interval? Round to two decimal places.

15. Suppose your sample size had been 30 rather than 20. What would the 95% confidence interval be then? Round to two decimal places

8.3: Confidence Interval for a Population Proportion

Use this information to answer the next four exercises. You conduct a poll of 500 randomly selected city residents, asking them if they own an automobile. 280 say they do own an automobile, and 220 say they do not.

16. Find the sample proportion and sample standard deviation for this data.

17. What is the 95% two-sided confidence interval? Round to four decimal places.

18. Calculate the 90% confidence interval. Round to four decimal places.

19. Calculate the 99% confidence interval. Round to four decimal places.

Use the following information to answer the next three exercises. You are planning to conduct a poll of community members age 65 and older, to determine how many own mobile phones. You want to produce an estimate whose 95% confidence interval will be within four percentage points (plus or minus) the true population proportion. Use an estimated population proportion of 0.5.

20. What sample size do you need?

21. Suppose you knew from prior research that the population proportion was 0.6. What sample size would you need?

22. Suppose you wanted a 95% confidence interval within three percentage points of the population. Assume the population proportion is 0.5. What sample size do you need?

9.1: Null and Alternate Hypotheses

23. In your state, 58 percent of registered voters in a community are registered as Republicans. You want to conduct a study to see if this also holds up in your community. State the null and alternative hypotheses to test this.

24. You believe that at least 58 percent of registered voters in a community are registered as Republicans. State the null and alternative hypotheses to test this.

25. The mean household value in a city is $268,000. You believe that the mean household value in a particular neighborhood is lower than the city average. Write the null and alternative hypotheses to test this.

26. State the appropriate alternative hypothesis to this null hypothesis: H_0: $\mu = 107$

27. State the appropriate alternative hypothesis to this null hypothesis: H_0: $p < 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. If you reject H_0 when H_0 is correct, what type of error is this?

29. If you fail to reject H_0 when H_0 is false, what type of error is this?

30. What is the relationship between the Type II error and the power of a test?

31. A new blood test is being developed to screen patients for cancer. Positive results are followed up by a more accurate (and expensive) test. It is assumed that the patient does not have cancer. Describe the null hypothesis, the Type I and Type II errors for this situation, and explain which type of error is more serious.

32. Explain in words what it means that a screening test for TB has an α level of 0.10. The null hypothesis is that the patient does not have TB.

33. Explain in words what it means that a screening test for TB has a β level of 0.20. The null hypothesis is that the patient does not have TB.

34. Explain in words what it means that a screening test for TB has a power of 0.80.

9.3: Distribution Needed for Hypothesis Testing

35. If you are conducting a hypothesis test of a single population mean, and you do not know the population variance, what test will you use if the sample size is 10 and the population is normal?

36. If you are conducting a hypothesis test of a single population mean, and you know the population variance, what test will you use?

37. If you are conducting a hypothesis test of a single population proportion, with *np* and *nq* greater than or equal to five, what test will you use, and with what parameters?

38. Published information indicates that, on average, college students spend less than 20 hours studying per week. You draw a sample of 25 students from your college, and find the sample mean to be 18.5 hours, with a standard deviation of 1.5 hours. What distribution will you use to test whether study habits at your college are the same as the national average, and why?

39. A published study says that 95 percent of American children are vaccinated against measles, with a standard deviation of 1.5 percent. You draw a sample of 100 children from your community and check their vaccination records, to see if the vaccination rate in your community is the same as the national average. What distribution will you use for this test, and why?

9.4: Rare Events, the Sample, Decision, and Conclusion

40. You are conducting a study with an α level of 0.05. If you get a result with a p-value of 0.07, what will be your decision?

41. You are conducting a study with $\alpha = 0.01$. If you get a result with a *p*-value of 0.006, what will be your decision?

Use the following information to answer the next five exercises. According to the World Health Organization, the average height of a one-year-old child is 29”. You believe children with a particular disease are smaller than

average, so you draw a sample of 20 children with this disease and find a mean height of 27.5” and a sample standard deviation of 1.5”.

42. What are the null and alternative hypotheses for this study?

43. What distribution will you use to test your hypothesis, and why?

44. What is the test statistic and the *p*-value?

45. Based on your sample results, what is your decision?

46. Suppose the mean for your sample was 25.0. Redo the calculations and describe what your decision would be.

9.5: Additional Information and Full Hypothesis Test Examples

47. You conduct a study using $\alpha = 0.05$. What is the level of significance for this study?

48. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu = 35.5$
H_a: $\mu \neq 35.5$
Will you conduct a one-tailed or two-tailed test?

49. You conduct a study, based on a sample drawn from a normally distributed population with a known variance, with the following hypotheses:
H_0: $\mu \geq 35.5$
H_a: $\mu < 35.5$
Will you conduct a one-tailed or two-tailed test?

Use the following information to answer the next three exercises. Nationally, 80 percent of adults own an automobile. You are interested in whether the same proportion in your community own cars. You draw a sample of 100 and find that 75 percent own cars.

50. What are the null and alternative hypotheses for this study?

51. What test will you use, and why?

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. You conduct a poll of political opinions, interviewing both members of 50 married couples. Are the groups in this study independent or matched?

53. You are testing a new drug to treat insomnia. You randomly assign 80 volunteer subjects to either the experimental (new drug) or control (standard treatment) conditions. Are the groups in this study independent or matched?

54. You are investigating the effectiveness of a new math textbook for high school students. You administer a pretest to a group of students at the beginning of the semester, and a posttest at the end of a year’s instruction using this textbook, and compare the results. Are the groups in this study independent or matched?

Use the following information to answer the next two exercises. You are conducting a study of the difference in time at two colleges for undergraduate degree completion. At College A, students take an average of 4.8 years to complete an undergraduate degree, while at College B, they take an average of 4.2 years. The pooled standard deviation for this data is 1.6 years

55. Calculate Cohen's *d* and interpret it.

56. Suppose the mean time to earn an undergraduate degree at College A was 5.2 years. Calculate the effect size and interpret it.

57. You conduct an independent-samples t-test with sample size ten in each of two groups. If you are conducting a two-tailed hypothesis test with $\alpha = 0.01$, what p-values will cause you to reject the null hypothesis?

58. You conduct an independent samples *t*-test with sample size 15 in each group, with the following hypotheses:
H_0: $\mu \geq 110$
H_a: $\mu < 110$
If $\alpha = 0.05$, what *t*-values will cause you to reject the null hypothesis?

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

Use the following information to answer the next six exercises. College students in the sciences often complain that they must spend more on textbooks each semester than students in the humanities. To test this, you draw random samples of 50 science and 50 humanities students from your college, and record how much each spent last semester on textbooks. Consider the science students to be group one, and the humanities students to be group two.

59. What is the random variable for this study?

60. What are the null and alternative hypotheses for this study?

61. If the 50 science students spent an average of $530 with a sample standard deviation of $20 and the 50 humanities students spent an average of $380 with a sample standard deviation of $15, would you not reject or reject the null hypothesis? Use an alpha level of 0.05. What is your conclusion?

62. What would be your decision, if you were using $\alpha = 0.01$?

10.3: Comparing Two Independent Population Proportions

Use the information to answer the next six exercises. You want to know if proportion of homes with cable television service differs between Community A and Community B. To test this, you draw a random sample of 100 for each and record whether they have cable service.

63. What are the null and alternative hypotheses for this study

64. If 65 households in Community A have cable service, and 78 households in community B, what is the pooled proportion?

65. At $\alpha = 0.03$, will you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

66. Using an alpha value of 0.01, would you reject the null hypothesis? What is your conclusion? 65 households in Community A have cable service, and 78 households in community B. 100 households in each community were surveyed.

10.4: Matched or Paired Samples

Use the following information to answer the next five exercises. You are interested in whether a particular exercise program helps people lose weight. You conduct a study in which you weigh the participants at the start of the study, and again at the conclusion, after they have participated in the exercise program for six months. You

compare the results using a matched-pairs t-test, in which the data is {weight at conclusion – weight at start}. You believe that, on average, the participants will have lost weight after six months on the exercise program.

67. What are the null and alternative hypotheses for this study?

68. Calculate the test statistic, assuming that x_d = –5, s_d = 6, and n = 30 (pairs).

69. What are the degrees of freedom for this statistic?

70. Using α = 0.05, what is your decision regarding the effectiveness of this program in causing weight loss? What is the conclusion?

71. What would it mean if the t-statistic had been 4.56, and what would have been your decision in that case?

11.1: Facts About the Chi-Square Distribution

72. What is the mean and standard deviation for a chi-square distribution with 20 degrees of freedom?

11.2: Goodness-of-Fit Test

Use the following information to answer the next four exercises. Nationally, about 66 percent of high school graduates enroll in higher education. You perform a chi-square goodness of fit test to see if this same proportion applies to your high school’s most recent graduating class of 200. Your null hypothesis is that the national distribution also applies to your high school.

73. What are the expected numbers of students from your high school graduating class enrolled and not enrolled in higher education?

74. Fill out the rest of this table.

	Observed (*O*)	**Expected (*E*)**	***O* – *E***	**(*O* – *E*)2**	$\frac{(O-E)^2}{z}$
Enrolled	145				
Not enrolled	55				

75. What are the degrees of freedom for this chi-square test?

76. What is the chi-square test statistic and the p-value. At the 5% significance level, what do you conclude?

77. For a chi-square distribution with 92 degrees of freedom, the curve ______________.

78. For a chi-square distribution with five degrees of freedom, the curve is _______________.

11.3: Test of Independence

Use the following information to answer the next four exercises. You are considering conducting a chi-square test of independence for the data in this table, which displays data about cell phone ownership for freshman and seniors at a high school. Your null hypothesis is that cell phone ownership is independent of class standing.

79. Compute the expected values for the cells.

	Cell = Yes	**Cell = No**
Freshman	100	150
Senior	200	50

80. Compute $\frac{(O-E)^2}{z}$ for each cell, where *O* = observed and *E* = expected.

81. What is the chi-square statistic and degrees of freedom for this study?

82. At the α = 0.5 significance level, what is your decision regarding the null hypothesis?

11.4: Test of Homogeneity

83. You conduct a chi-square test of homogeneity for data in a five by two table. What is the degrees of freedom for this test?

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. A 2013 poll in the State of California surveyed people about taxing sugar-sweetened beverages. The results are presented in the following table, and are classified by ethnic group and response type. Are the poll responses independent of the participants’ ethnic group? Conduct a hypothesis test at the 5% significance level.

Ethnic Group \ Response Type	**Favor**	**Oppose**	**No Opinion**	**Row Total**
White / Non-Hispanic	234	433	43	710
Latino	147	106	19	272
African American	24	41	6	71
Asian American	54	48	16	118
Column Total	459	628	84	1171

85. In a test of homogeneity, what must be true about the expected value of each cell?

86. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of independence?

87. Stated in general terms, what are the null and alternative hypotheses for the chi-square test of homogeneity?

11.6: Test of a Single Variance

88. A lab test claims to have a variance of no more than five. You believe the variance is greater. What are the null and alternative hypothesis to test this?

Practice Test 3 Solutions

8.1: Confidence Interval, Single Population Mean, Population Standard Deviation Known, Normal

1. $\frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{30}} = 0.73$

2. normal

3. 0.025 or 2.5%; A 95% confidence interval contains 95% of the probability, and excludes five percent, and the five percent excluded is split evenly between the upper and lower tails of the distribution.

4. *z*-score = 1.96; $EBM = \ z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)\,(0.73) = \ 1.4308$

5. 41 ± 1.43 = (39.57, 42.43); Using the calculator function Zinterval, answer is (40.74, 41.26. Answers differ due to rounding.

6. The *z*-value for a 90% confidence interval is 1.645, so *EBM* = 1.645(0.73) = 1.20085.
The 90% confidence interval is 41 ± 1.20 = (39.80, 42.20).
The calculator function Zinterval answer is (40.78, 41.23). Answers differ due to rounding.

7. The standard error of measurement is: $\frac{\sigma}{\sqrt{n}} = \ \frac{4}{\sqrt{50}} = 0.57$
$EBM = \ z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) = (1.96)\,(0.57) = \ 1.12$
The 95% confidence interval is 41 ± 1.12 = (39.88, 42.12).
The calculator function Zinterval answer is (40.84, 41.16). Answers differ due to rounding.

8. The 99% confidence interval, because it includes all but one percent of the distribution. The 95% confidence interval will be narrower, because it excludes five percent of the distribution.

8.2: Confidence Interval, Single Population Mean, Standard Deviation Unknown, Student's t

9. The *t*-distribution will have more probability in its tails ("thicker tails") and less probability near the mean of the distribution ("shorter in the center").

10. Both distributions are symmetrical and centered at zero.

11. *df* = *n* – 1 = 20 – 1 = 19

12. You can get the *t*-value from a probability table or a calculator. In this case, for a *t*-distribution with 19 degrees of freedom, and a 95% two-sided confidence interval, the value is 2.093, i.e.,
$t_{\frac{\alpha}{2}} = 2.093$. The calculator function is invT(0.975, 19).

13. $EBM = \ t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = (2.093) \left(\frac{0.3}{\sqrt{20}} \right) = \ 0.140$
98.4 ± 0.14 = (98.26, 98.54).
The calculator function Tinterval answer is (98.26, 98.54).

14. $t_{\frac{\alpha}{2}} = 2.861$. The calculator function is invT(0.995, 19).

$$EBM = t_{\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right) = (2.861) \left(\frac{0.3}{\sqrt{20}} \right) = 0.192$$

98.4 ± 0.19 = (98.21, 98.59). The calculator function Tinterval answer is (98.21, 98.59).

15. *df* = *n* – 1 = 30 – 1 = 29. $t_{\frac{\alpha}{2}} \ = 2.045$
$EBM = \ z_t \left(\frac{s}{\sqrt{n}} \right) = (2.045) \left(\frac{0.3}{\sqrt{30}} \right) = \ 0.112$
98.4 ± 0.11 = (98.29, 98.51). The calculator function Tinterval answer is (98.29, 98.51).

8.3: Confidence Interval for a Population Proportion

16. $p' = \frac{280}{500} = 0.56$
$q' = 1 - p' = 1 - 0.56 = 0.44$
$s = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.56(0.44)}{500}} = 0.0222$

17. Because you are using the normal approximation to the binomial, $z_{\frac{\alpha}{2}} = 1.96$.
Calculate the error bound for the population (*EBP*):
$EBP = \ z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} = 1.96\,(0.222) = 0.0435$
Calculate the 95% confidence interval:
0.56 ± 0.0435 = (0.5165, 0.6035).
The calculator function 1-PropZint answer is (0.5165, 0.6035).

18. $z_{\frac{\alpha}{2}} = 1.64$
$EBP = \ z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} = 1.64\,(0.0222) = 0.0364$
0.56 ± 0.03 = (0.5236, 0.5964). The calculator function 1-PropZint answer is (0.5235, 0.5965)

19. $z_{\frac{\alpha}{2}} = 2.58$
$EBP = \ z_{\frac{\alpha}{2}} \sqrt{\frac{pq}{n}} = 2.58\,(0.0222) = \ 0.0573$
0.56 ± 0.05 = (0.5127, 0.6173).
The calculator function 1-PropZint answer is (0.5028, 0.6172).

20. *EBP* = 0.04 (because 4% = 0.04)
$z_{\frac{\alpha}{2}} = 1.96$ for a 95% confidence interval
$n = \ \frac{z^2 pq}{EBP^2} = \ \frac{1.96^2(0.5)(0.5)}{0.04^2} \ = \ \frac{0.9604}{0.0016} = 600.25$
You need 601 subjects (rounding upward from 600.25).

21. $n = \ \frac{n^2 pq}{EBP^2} \ = \ \frac{1.96^2(0.6)(0.4)}{0.04^2} \ = \ \frac{0.9220}{0.0016} = 576.24$
You need 577 subjects (rounding upward from 576.24).

22. $n = \ \frac{n^2 pq}{EBP^2} \ = \ \frac{1.96^2(0.5)(0.5)}{0.03^2} \ = \ \frac{0.9604}{0.0009} = 1067.11$
You need 1,068 subjects (rounding upward from 1,067.11).

9.1: Null and Alternate Hypotheses

23. H_0: $p = 0.58$
H_a: $p \neq 0.58$

24. H_0: $p \geq 0.58$
H_a: $p < 0.58$

25. H_0: $\mu \geq \$268,000$
H_a: $\mu < \$268,000$

26. H_a: $\mu \neq 107$

27. H_a: $p \geq 0.25$

9.2: Outcomes and the Type I and Type II Errors

28. a Type I error

29. a Type II error

30. Power = $1 – \beta = 1 – P$(Type II error).

31. The null hypothesis is that the patient does not have cancer. A Type I error would be detecting cancer when it is not present. A Type II error would be not detecting cancer when it is present. A Type II error is more serious, because failure to detect cancer could keep a patient from receiving appropriate treatment.

32. The screening test has a ten percent probability of a Type I error, meaning that ten percent of the time, it will detect TB when it is not present.

33. The screening test has a 20 percent probability of a Type II error, meaning that 20 percent of the time, it will fail to detect TB when it is in fact present.

34. Eighty percent of the time, the screening test will detect TB when it is actually present.

9.3: Distribution Needed for Hypothesis Testing

35. The Student's *t*-test.

36. The normal distribution or *z*-test.

37. The normal distribution with $\mu = p$ and $\sigma = \sqrt{\frac{pq}{n}}$

38. t_{24}. You use the *t*-distribution because you don't know the population standard deviation, and the degrees of freedom are 24 because $df = n – 1$.

39. $X \sim N\left(0.95, \frac{0.051}{\sqrt{100}}\right)$
Because you know the population standard deviation, and have a large sample, you can use the normal distribution.

9.4: Rare Events, the Sample, Decision, and Conclusion

40. Fail to reject the null hypothesis, because $\alpha \leq p$

41. Reject the null hypothesis, because $\alpha \geq p$.

42. H_0: $\mu \geq 29.0$”
H_a: $\mu < 29.0$”

43. t_{19}. Because you do not know the population standard deviation, use the *t*-distribution. The degrees of freedom are 19, because $df = n - 1$.

44. The test statistic is –4.4721 and the *p*-value is 0.00013 using the calculator function TTEST.

45. With $\alpha = 0.05$, reject the null hypothesis.

46. With $\alpha = 0.05$, the *p*-value is almost zero using the calculator function TTEST so reject the null hypothesis.

9.5: Additional Information and Full Hypothesis Test Examples

47. The level of significance is five percent.

48. two-tailed

49. one-tailed

50. H_0: $p = 0.8$
H_a: $p \neq 0.8$

51. You will use the normal test for a single population proportion because *np* and *nq* are both greater than five.

10.1: Comparing Two Independent Population Means with Unknown Population Standard Deviations

52. They are matched (paired), because you interviewed married couples.

53. They are independent, because participants were assigned at random to the groups.

54. They are matched (paired), because you collected data twice from each individual.

55. $d = \frac{x_1 - x_2}{s_{pooled}} = \frac{4.8 - 4.2}{1.6} = 0.375$
This is a small effect size, because 0.375 falls between Cohen’s small (0.2) and medium (0.5) effect sizes.

56. $d = \frac{x_1 - x_2}{s_{pooled}} = \frac{5.2 - 4.2}{1.6} = 0.625$
The effect size is 0.625. By Cohen’s standard, this is a medium effect size, because it falls between the medium (0.5) and large (0.8) effect sizes.

57. *p*-value < 0.01.

58. You will only reject the null hypothesis if you get a value significantly below the hypothesized mean of 110.

10.2: Comparing Two Independent Population Means with Known Population Standard Deviations

59. $X_1 - X_2$, i.e., the mean difference in amount spent on textbooks for the two groups.

60. H_0: $X_1 - X_2 \leq 0$
H_a: $X_1 - X_2 > 0$
This could also be written as:
H_0: $X_1 \leq X_2$
H_a: $X_1 > X_2$

61. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

62. Using the calculator function 2-SampTtest, reject the null hypothesis. At the 1% significance level, there is sufficient evidence to conclude that the science students spend more on textbooks than the humanities students.

10.3: Comparing Two Independent Population Proportions

63. H_0: $p_A = p_B$
H_a: $p_A \neq p_B$

64. $p_c = \frac{x_A + x_A}{n_A + n_A} = \frac{65+78}{100+100} = 0.715$

65. Using the calculator function 2-PropZTest, the p-value = 0.0417. Reject the null hypothesis. At the 3% significance level, here is sufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

66. Using the calculator function 2-PropZTest, the *p*-value = 0.0417. Do not reject the null hypothesis. At the 1% significance level, there is insufficient evidence to conclude that there is a difference between the proportions of households in the two communities that have cable service.

10.4: Matched or Paired Samples

67. H_0: $x_d \geq 0$
H_a: $x_d < 0$

68. $t = -4.5644$

69. $df = 30 - 1 = 29$.

70. Using the calculator function TTEST, the *p*-value = 0.00004 so reject the null hypothesis. At the 5% level, there is sufficient evidence to conclude that the participants lost weight, on average.

71. A positive *t*-statistic would mean that participants, on average, gained weight over the six months.

11.1: Facts About the Chi-Square Distribution

72. $\mu = df = 20$
$\sigma = \sqrt{2(df)} = \sqrt{40} = 6.32$

11.2: Goodness-of-Fit Test

73. Enrolled = 200(0.66) = 132. Not enrolled = 200(0.34) = 68

74.

	Observed (O)	**Expected (E)**	**O – E**	**(O – E)2**	$\frac{(O-E)^2}{z}$
Enrolled	145	132	145 – 132 = 13	169	$\frac{169}{132} = 1.280$
Not enrolled	55	68	55 – 68 = –13	169	$\frac{169}{68} = 2.485$

75. *df* = *n* – 1 = 2 – 1 = 1.

76. Using the calculator function Chi-square GOF – Test (in STAT TESTS), the test statistic is 3.7656 and the p-value is 0.0523. Do not reject the null hypothesis. At the 5% significance level, there is insufficient evidence to conclude that high school most recent graduating class distribution of enrolled and not enrolled does not fit that of the national distribution.

77. approximates the normal

78. skewed right

11.3: Test of Independence

79.

	Cell = Yes	**Cell = No**	**Total**
Freshman	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Senior	$\frac{250(300)}{500} = 150$	$\frac{250(200)}{500} = 100$	250
Total	300	200	500

80. $\frac{(100-150)^2}{150} = 16.67$
$\frac{(150-100)^2}{100} = 25$
$\frac{(200-100)^2}{150} = 16.67$
$\frac{(50-100)^2}{100} = 25$

81. Chi-square = 16.67 + 25 + 16.67 + 25 = 83.34.
df = (*r* – 1)(*c* – 1) = 1

82. p-value = P(Chi-square, 83.34) = 0
Reject the null hypothesis.
You could also use the calculator function STAT TESTS Chi-Square – Test.

11.4: Test of Homogeneity

83. The table has five rows and two columns. $df = (r - 1)(c - 1) = (4)(1) = 4$.

11.5: Comparison Summary of the Chi-Square Tests: Goodness-of-Fit, Independence and Homogeneity

84. Using the calculator function (STAT TESTS) Chi-square Test, the p-value = 0. Reject the null hypothesis. At the 5% significance level, there is sufficient evidence to conclude that the poll responses independent of the participants' ethnic group.

85. The expected value of each cell must be at least five.

86. H_0: The variables are independent.
H_a: The variables are not independent.

87. H_0: The populations have the same distribution.
H_a: The populations do not have the same distribution.

11.6: Test of a Single Variance

88. H_0: $\sigma^2 \leq 5$
H_a: $\sigma^2 > 5$

Practice Test 4

12.1 Linear Equations

1. Which of the following equations is/are linear?

a. $y = -3x$
b. $y = 0.2 + 0.74x$
c. $y = -9.4 - 2x$
d. A and B
e. A, B, and C

2. To complete a painting job requires four hours setup time plus one hour per 1,000 square feet. How would you express this information in a linear equation?

3. A statistics instructor is paid a per-class fee of $2,000 plus $100 for each student in the class. How would you express this information in a linear equation?

4. A tutoring school requires students to pay a one-time enrollment fee of $500 plus tuition of $3,000 per year. Express this information in an equation.

12.2: Slope and *Y*-intercept of a Linear Equation

Use the following information to answer the next four exercises. For the labor costs of doing repairs, an auto mechanic charges a flat fee of $75 per car, plus an hourly rate of $55.

5. What are the independent and dependent variables for this situation?

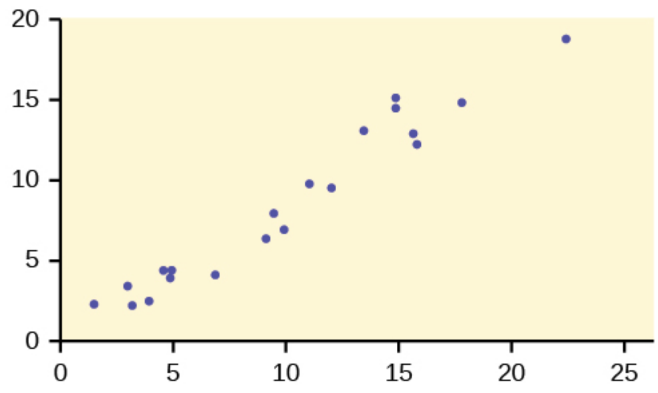
6. Write the equation and identify the slope and intercept.

7. What is the labor charge for a job that takes 3.5 hours to complete?

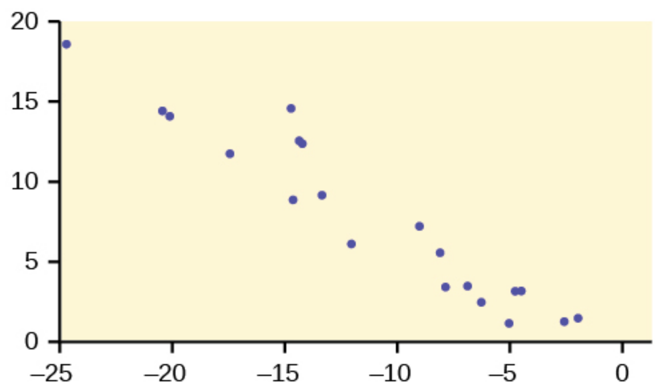
8. One job takes 2.4 hours to complete, while another takes 6.3 hours. What is the difference in labor costs for these two jobs?

12.3: Scatter Plots

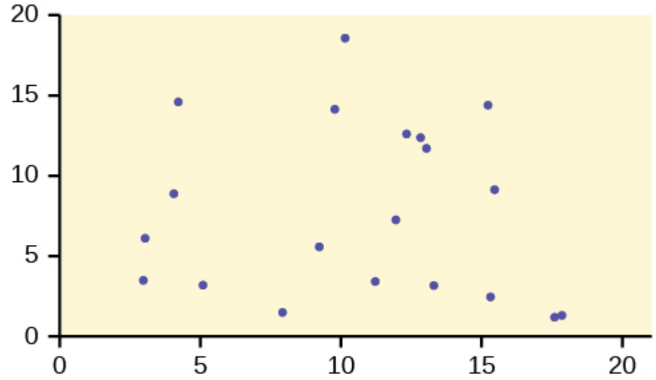
9. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



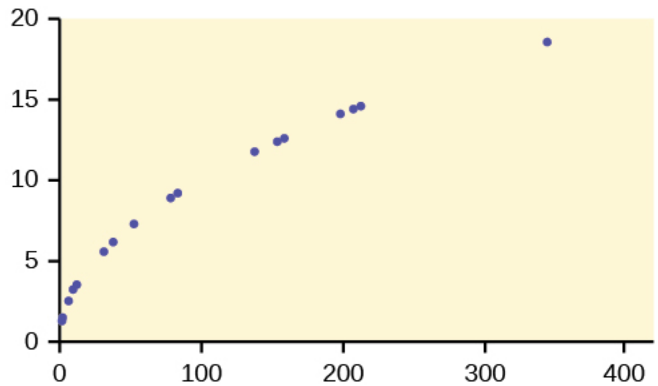
10. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



11. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



12. Describe the pattern in this scatter plot, and decide whether the *X* and *Y* variables would be good candidates for linear regression.



12.4: The Regression Equation

Use the following information to answer the next four exercises. Height (in inches) and weight (In pounds) in a sample of college freshman men have a linear relationship with the following summary statistics:
x = 68.4
y =141.6
s_x = 4.0
s_y = 9.6
r = 0.73
Let Y = weight and X = height, and write the regression equation in the form:
$\hat{y} = a + bx$

13. What is the value of the slope?

14. What is the value of the *y* intercept?

15. Write the regression equation predicting weight from height in this data set, and calculate the predicted weight for someone 68 inches tall.

12.5: Correlation Coefficient and Coefficient of Determination

16. The correlation between body weight and fuel efficiency (measured as miles per gallon) for a sample of 2,012 model cars is –0.56. Calculate the coefficient of determination for this data and explain what it means.

17. The correlation between high school GPA and freshman college GPA for a sample of 200 university students is 0.32. How much variation in freshman college GPA is not explained by high school GPA?

18. Rounded to two decimal places what correlation between two variables is necessary to have a coefficient of determination of at least 0.50?

12.6: Testing the Significance of the Correlation Coefficient

19. Write the null and alternative hypotheses for a study to determine if two variables are significantly correlated.

20. In a sample of 30 cases, two variables have a correlation of 0.33. Do a *t*-test to see if this result is significant at the $\alpha = 0.05$ level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

21. In a sample of 25 cases, two variables have a correlation of 0.45. Do a *t*-test to see if this result is significant at the $\alpha = 0.05$ level. Use the formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

12.7: Prediction

Use the following information to answer the next two exercises. A study relating the grams of potassium (*Y*) to the grams of fiber (*X*) per serving in enriched flour products (bread, rolls, etc.) produced the equation:

$$\hat{y} = 25 + 16x$$

22. For a product with five grams of fiber per serving, what are the expected grams of potassium per serving?

23. Comparing two products, one with three grams of fiber per serving and one with six grams of fiber per serving, what is the expected difference in grams of potassium per serving?

12.8: Outliers

24. In the context of regression analysis, what is the definition of an outlier, and what is a rule of thumb to evaluate if a given value in a data set is an outlier?

25. In the context of regression analysis, what is the definition of an influential point, and how does an influential point differ from an outlier?

26. The least squares regression line for a data set is $\hat{y} = 5 + 0.3x$ and the standard deviation of the residuals is 0.4. Does a case with the values $x = 2$, $y = 6.2$ qualify as an outlier?

27. The least squares regression line for a data set is $\hat{y} = 2.3 - 0.1x$ and the standard deviation of the residuals is 0.13. Does a case with the values $x = 4.1$, $y = 2.34$ qualify as an outlier?

13.1: One-Way ANOVA

28. What are the five basic assumptions to be met if you want to do a one-way ANOVA?

29. You are conducting a one-way ANOVA comparing the effectiveness of four drugs in lowering blood pressure in hypertensive patients. What are the null and alternative hypotheses for this study?

30. What is the primary difference between the independent samples *t*-test and one-way ANOVA?

31. You are comparing the results of three methods of teaching geometry to high school students. The final exam scores X_1, X_2, X_3, for the samples taught by the different methods have the following distributions:
$X_1 \sim N(85, 3.6)$
$X_1 \sim N(82, 4.8)$
$X_1 \sim N(79, 2.9)$
Each sample includes 100 students, and the final exam scores have a range of 0–100. Assuming the samples are independent and randomly selected, have the requirements for conducting a one-way ANOVA been met? Explain why or why not for each assumption.

32. You conduct a study comparing the effectiveness of four types of fertilizer to increase crop yield on wheat farms. When examining the sample results, you find that two of the samples have an approximately normal distribution, and two have an approximately uniform distribution. Is this a violation of the assumptions for conducting a one-way ANOVA?

13.2: The *F* Distribution

Use the following information to answer the next seven exercises. You are conducting a study of three types of feed supplements for cattle to test their effectiveness in producing weight gain among calves whose feed includes one of the supplements. You have four groups of 30 calves (one is a control group receiving the usual feed, but no supplement). You will conduct a one-way ANOVA after one year to see if there are difference in the mean weight for the four groups.

33. What is SS_{within} in this experiment, and what does it mean?

34. What is $SS_{between}$ in this experiment, and what does it mean?

35. What are *k* and *i* for this experiment?

36. If $SS_{within} = 374.5$ and $SS_{total} = 621.4$ for this data, what is $SS_{between}$?

37. What are $MS_{between}$, and MS_{within}, for this experiment?

38. What is the *F* Statistic for this data?

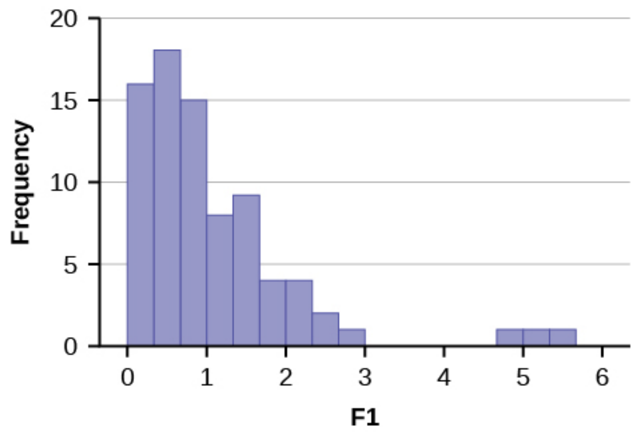
39. If there had been 35 calves in each group, instead of 30, with the sums of squares remaining the same, would the *F* Statistic be larger or smaller?

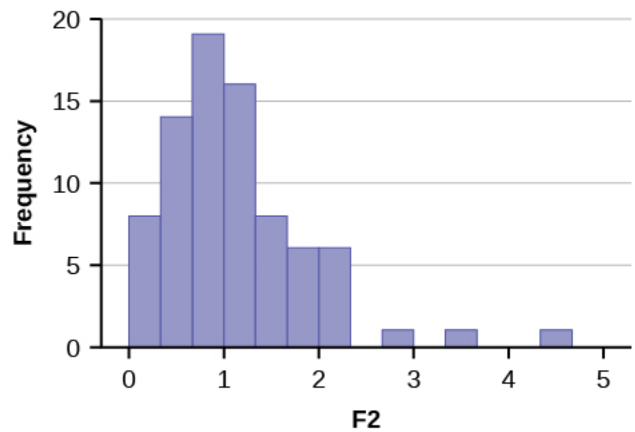
13.3: Facts About the *F* Distribution

40. Which of the following numbers are possible *F* Statistics?

a. 2.47
b. 5.95
c. –3.61
d. 7.28
e. 0.97

41. Histograms *F*1 and *F*2 below display the distribution of cases from samples from two populations, one distributed $F_{3,15}$ and one distributed $F_{5,500}$. Which sample came from which population?





42. The F Statistic from an experiment with $k = 3$ and $n = 50$ is 3.67. At $\alpha = 0.05$, will you reject the null hypothesis?

43. The F Statistic from an experiment with $k = 4$ and $n = 100$ is 4.72. At $\alpha = 0.01$, will you reject the null hypothesis?

13.4: Test of Two Variances

44. What assumptions must be met to perform the F test of two variances?

45. You believe there is greater variance in grades given by the math department at your university than in the English department. You collect all the grades for undergraduate classes in the two departments for a semester, and compute the variance of each, and conduct an F test of two variances. What are the null and alternative hypotheses for this study?

Practice Test 4 Solutions

12.1 Linear Equations

1. e. A, B, and C.
All three are linear equations of the form $y = mx + b$.

2. Let y = the total number of hours required, and x the square footage, measured in units of 1,000. The equation is: $y = x + 4$

3. Let y = the total payment, and x the number of students in a class. The equation is: $y = 100(x) + 2{,}000$

4. Let y = the total cost of attendance, and x the number of years enrolled. The equation is: $y = 3{,}000(x) + 500$

12.2: Slope and Y-intercept of a Linear Equation

5. The independent variable is the hours worked on a car. The dependent variable is the total labor charges to fix a car.

6. Let y = the total charge, and x the number of hours required. The equation is: $y = 55x + 75$
The slope is 55 and the intercept is 75.

7. $y = 55(3.5) + 75 = 267.50$

8. Because the intercept is included in both equations, while you are only interested in the difference in costs, you do not need to include the intercept in the solution. The difference in number of hours required is: 6.3 – 2.4 = 3.9. Multiply this difference by the cost per hour: 55(3.9) = 214.5.
The difference in cost between the two jobs is $214.50.

12.3: Scatter Plots

9. The X and Y variables have a strong linear relationship. These variables would be good candidates for analysis with linear regression.

10. The X and Y variables have a strong negative linear relationship. These variables would be good candidates for analysis with linear regression.

11. There is no clear linear relationship between the X and Y variables, so they are not good candidates for linear regression.

12. The X and Y variables have a strong positive relationship, but it is curvilinear rather than linear. These variables are not good candidates for linear regression.

12.4: The Regression Equation

13. $r\left(\frac{s_y}{s_x}\right) = 0.73\left(\frac{9.6}{4.0}\right) = 1.752 \approx 1.75$

14. $a = y - bx = 141.6 - 1.752(68.4) = 21.7632 \approx 21.76$

15. $\hat{y} = 21.76 + 1.75(68) = 140.76$

12.5: Correlation Coefficient and Coefficient of Determination

16. The coefficient of determination is the square of the correlation, or r^2.
For this data, $r^2 = (–0.56)2 = 0.3136 \approx 0.31$ or 31%. This means that 31 percent of the variation in fuel efficiency can be explained by the bodyweight of the automobile.

17. The coefficient of determination = $0.32^2 = 0.1024$. This is the amount of variation in freshman college GPA that can be explained by high school GPA. The amount that cannot be explained is 1 – 0.1024 = 0.8976 ≈ 0.90. So about 90 percent of variance in freshman college GPA in this data is not explained by high school GPA.

18. $r = \sqrt{r^2}$
$\sqrt{0.5} = 0.707106781 \approx 0.71$
You need a correlation of 0.71 or higher to have a coefficient of determination of at least 0.5.

12.6: Testing the Significance of the Correlation Coefficient

19. H_0: ρ = 0
H_a: ρ ≠ 0

20. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.33\sqrt{30-2}}{\sqrt{1-0.33^2}} = 1.85$
The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{29} distribution is 2.045. Your value is less than this, so you fail to reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.
Using the calculator function tcdf, the *p*-value is 2tcdf(1.85, 10^99, 29) = 0.0373. Do not reject the null hypothesis and conclude that the study produced no evidence that the variables are significantly correlated.

21. $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.45\sqrt{25-2}}{\sqrt{1-0.45^2}} = 2.417$
The critical value for $\alpha = 0.05$ for a two-tailed test using the t_{24} distribution is 2.064. Your value is greater than this, so you reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.
Using the calculator function tcdf, the p-value is 2tcdf(2.417, 10^99, 24) = 0.0118. Reject the null hypothesis and conclude that the study produced evidence that the variables are significantly correlated.

12.7: Prediction

22. $\hat{y} = 25 + 16(5) = 105$

23. Because the intercept appears in both predicted values, you can ignore it in calculating a predicted difference score. The difference in grams of fiber per serving is 6 – 3 = 3 and the predicted difference in grams of potassium per serving is (16)(3) = 48.

12.8: Outliers

24. An outlier is an observed value that is far from the least squares regression line. A rule of thumb is that a point more than two standard deviations of the residuals from its predicted value on the least squares regression line is an outlier.

25. An influential point is an observed value in a data set that is far from other points in the data set, in a horizontal direction. Unlike an outlier, an influential point is determined by its relationship with other values in the data set, not by its relationship to the regression line.

26. The predicted value for y is: $\hat{y} = 5 + 0.3x = 5.6$. The value of 6.2 is less than two standard deviations from the predicted value, so it does not qualify as an outlier.
Residual for (2, 6.2): 6.2 – 5.6 = 0.6 (0.6 < 2(0.4))

27. The predicted value for y is: $\hat{y}$ = 2.3 – 0.1(4.1) = 1.89. The value of 2.32 is more than two standard deviations from the predicted value, so it qualifies as an outlier.
Residual for (4.1, 2.34): 2.32 – 1.89 = 0.43 (0.43 > 2(0.13))

13.1: One-Way ANOVA

28.

1. Each sample is drawn from a normally distributed population
2. All samples are independent and randomly selected.
3. The populations from which the samples are draw have equal standard deviations.
4. The factor is a categorical variable.
5. The response is a numerical variable.

29. H_0: $\mu 1 = \mu 2 = \mu 3 = \mu 4$
H_a: At least two of the group means $\mu 1$, $\mu 2$, $\mu 3$, $\mu 4$ are not equal.

30. The independent samples *t*-test can only compare means from two groups, while one-way ANOVA can compare means of more than two groups.

31. Each sample appears to have been drawn from a normally distributed populations, the factor is a categorical variable (method), the outcome is a numerical variable (test score), and you were told the samples were independent and randomly selected, so those requirements are met. However, each sample has a different standard deviation, and this suggests that the populations from which they were drawn also have different standard deviations, which is a violation of an assumption for one-way ANOVA. Further statistical testing will be necessary to test the assumption of equal variance before proceeding with the analysis.

32. One of the assumptions for a one-way ANOVA is that the samples are drawn from normally distributed populations. Since two of your samples have an approximately uniform distribution, this casts doubt on whether this assumption has been met. Further statistical testing will be necessary to determine if you can proceed with the analysis.

13.2: The *F* Distribution

33. SS_{within} is the sum of squares within groups, representing the variation in outcome that cannot be attributed to the different feed supplements, but due to individual or chance factors among the calves in each group.

34. $SS_{between}$ is the sum of squares between groups, representing the variation in outcome that can be attributed to the different feed supplements.

35. k = the number of groups = 4
n_1 = the number of cases in group 1 = 30
n = the total number of cases = 4(30) = 120

36. $SS_{total} = SS_{within} + SS_{between}$ so $SS_{between} = SS_{total} - SS_{within}$
621.4 – 374.5 = 246.9

37. The mean squares in an ANOVA are found by dividing each sum of squares by its respective degrees of freedom (*df*).
For SS_{total}, $df = n - 1 = 120 - 1 = 119$.
For $SS_{between}$, df = k – 1 = 4 – 1 = 3.
For SS_{within}, df = 120 – 4 = 116.
$MS_{between} = \frac{246.9}{3} = 82.3$
$MS_{within} = \frac{374.5}{116} = 3.23$

38. $F = \frac{MS_{between}}{MS_{within}} = \frac{82.3}{3.23} = 25.48$

39. It would be larger, because you would be dividing by a smaller number. The value of $MS_{between}$ would not change with a change of sample size, but the value of MS_{within} would be smaller, because you would be dividing by a larger number (df_{within} would be 136, not 116). Dividing a constant by a smaller number produces a larger result.

13.3: Facts About the *F* Distribution

40. All but choice c, –3.61. *F* Statistics are always greater than or equal to 0.

41. As the degrees of freedom increase in an *F* distribution, the distribution becomes more nearly normal. Histogram *F*2 is closer to a normal distribution than histogram *F*1, so the sample displayed in histogram *F*1 was drawn from the $F_{3,15}$ population, and the sample displayed in histogram *F*2 was drawn from the $F_{5,500}$ population.

42. Using the calculator function Fcdf, *p*-value = Fcdf(3.67, 1E, 3,50) = 0.0182. Reject the null hypothesis.

43. Using the calculator function Fcdf, *p*-value = Fcdf(4.72, 1E, 4, 100) = 0.0016 Reject the null hypothesis.

13.4: Test of Two Variances

44. The samples must be drawn from populations that are normally distributed, and must be drawn from independent populations.

45. Let σ_M^2 = variance in math grades, and σ_E^2 = variance in English grades.
H_0: $\sigma_M^2 \leq \sigma_E^2$
H_a: $\sigma_M^2 > \sigma_E^2$

Practice Final Exam 1

Use the following information to answer the next two exercises: An experiment consists of tossing two, 12-sided dice (the numbers 1–12 are printed on the sides of each die).

- Let Event *A* = both dice show an even number.
- Let Event *B* = both dice show a number more than eight

1. Events *A* and *B* are:

a. mutually exclusive.
b. independent.
c. mutually exclusive and independent.
d. neither mutually exclusive nor independent.

2. Find *P*(*A*|*B*).

a. $\frac{2}{4}$
b. $\frac{16}{144}$
c. $\frac{4}{16}$
d. $\frac{2}{144}$

3. Which of the following are TRUE when we perform a hypothesis test on matched or paired samples?

a. Sample sizes are almost never small.
b. Two measurements are drawn from the same pair of individuals or objects.

c. Two sample means are compared to each other.
d. Answer choices b and c are both true.

Use the following information to answer the next two exercises: One hundred eighteen students were asked what type of color their bedrooms were painted: light colors, dark colors, or vibrant colors. The results were tabulated according to gender.

	Light colors	**Dark colors**	**Vibrant colors**
Female	20	22	28
Male	10	30	8

4. Find the probability that a randomly chosen student is male or has a bedroom painted with light colors.

a. $\frac{10}{118}$
b. $\frac{68}{118}$
c. $\frac{48}{118}$
d. $\frac{10}{48}$

5. Find the probability that a randomly chosen student is male given the student's bedroom is painted with dark colors.

a. $\frac{30}{118}$
b. $\frac{30}{48}$
c. $\frac{22}{118}$
d. $\frac{30}{52}$

Use the following information to answer the next two exercises: We are interested in the number of times a teenager must be reminded to do his or her chores each week. A survey of 40 mothers was conducted. [link] shows the results of the survey.

x	***P* (*x*)**
0	$\frac{2}{40}$
1	$\frac{5}{40}$
2	
3	$\frac{14}{40}$

x	$P(x)$
4	$\frac{7}{40}$
5	$\frac{4}{40}$

6. Find the probability that a teenager is reminded two times.

a. 8
b. $\frac{8}{40}$
c. $\frac{6}{40}$
d. 2

7. Find the expected number of times a teenager is reminded to do his or her chores.

a. 15
b. 2.78
c. 1.0
d. 3.13

Use the following information to answer the next two exercises: On any given day, approximately 37.5% of the cars parked in the De Anza parking garage are parked crookedly. We randomly survey 22 cars. We are interested in the number of cars that are parked crookedly.

8. For every 22 cars, how many would you expect to be parked crookedly, on average?

a. 8.25
b. 11
c. 18
d. 7.5

9. What is the probability that at least ten of the 22 cars are parked crookedly.

a. 0.1263
b. 0.1607
c. 0.2870
d. 0.8393

10. Using a sample of 15 Stanford-Binet IQ scores, we wish to conduct a hypothesis test. Our claim is that the mean IQ score on the Stanford-Binet IQ test is more than 100. It is known that the standard deviation of all Stanford-Binet IQ scores is 15 points. The correct distribution to use for the hypothesis test is:

a. Binomial
b. Student's *t*
c. Normal
d. Uniform

Use the following information to answer the next three exercises: De Anza College keeps statistics on the pass rate of students who enroll in math classes. In a sample of 1,795 students enrolled in Math 1A (1st quarter calculus), 1,428 passed the course. In a sample of 856 students enrolled in Math 1B (2nd quarter calculus), 662 passed. In general, are the pass rates of Math 1A and Math 1B statistically the same? Let A = the subscript for Math 1A and B = the subscript for Math 1B.

11. If you were to conduct an appropriate hypothesis test, the alternate hypothesis would be:

a. H_a: $p_A = p_B$
b. H_a: $p_A > p_B$
c. H_o: $p_A = p_B$
d. H_a: $p_A \neq p_B$

12. The Type I error is to:

a. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, the pass rates are different.
b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.
c. conclude that the pass rate for Math 1A is greater than the pass rate for Math 1B when, in fact, the pass rate for Math 1A is less than the pass rate for Math 1B.
d. conclude that the pass rate for Math 1A is the same as the pass rate for Math 1B when, in fact, they are the same.

13. The correct decision is to:

a. reject H_0
b. not reject H_0
c. There is not enough information given to conduct the hypothesis test

Kia, Alejandra, and Iris are runners on the track teams at three different schools. Their running times, in minutes, and the statistics for the track teams at their respective schools, for a one mile run, are given in the table below:

	Running Time	**School Average Running Time**	**School Standard Deviation**
Kia	4.9	5.2	0.15
Alejandra	4.2	4.6	0.25
Iris	4.5	4.9	0.12

14. Which student is the BEST when compared to the other runners at her school?

a. Kia
b. Alejandra
c. Iris
d. Impossible to determine

Use the following information to answer the next two exercises: The following adult ski sweater prices are from the Gorsuch Ltd. Winter catalog: \$212, \$292, \$278, \$199, \$280, \$236

Assume the underlying sweater price population is approximately normal. The null hypothesis is that the mean price of adult ski sweaters from Gorsuch Ltd. is at least \$275.

15. The correct distribution to use for the hypothesis test is:

a. Normal
b. Binomial

c. Student's *t*
d. Exponential

16. The hypothesis test:

a. is two-tailed.
b. is left-tailed.
c. is right-tailed.
d. has no tails.

17. Sara, a statistics student, wanted to determine the mean number of books that college professors have in their office. She randomly selected two buildings on campus and asked each professor in the selected buildings how many books are in his or her office. Sara surveyed 25 professors. The type of sampling selected is

a. simple random sampling.
b. systematic sampling.
c. cluster sampling.
d. stratified sampling.

18. A clothing store would use which measure of the center of data when placing orders for the typical "middle" customer?

a. mean
b. median
c. mode
d. IQR

19. In a hypothesis test, the *p*-value is

a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.
b. called the preconceived alpha.
c. compared to beta to decide whether to reject or not reject the null hypothesis.
d. Answer choices A and B are both true.

Use the following information to answer the next three exercises: A community college offers classes 6 days a week: Monday through Saturday. Maria conducted a study of the students in her classes to determine how many days per week the students who are in her classes come to campus for classes. In each of her 5 classes she randomly selected 10 students and asked them how many days they come to campus for classes. Each of her classes are the same size. The results of her survey are summarized in [link].

Number of Days on Campus	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
1	2		
2	12	.24	
3	10	.20	
4			.98
5	0		

Number of Days on Campus	Frequency	Relative Frequency	Cumulative Relative Frequency
6	1	.02	1.00

20. Combined with convenience sampling, what other sampling technique did Maria use?

a. simple random
b. systematic
c. cluster
d. stratified

21. How many students come to campus for classes four days a week?

a. 49
b. 25
c. 30
d. 13

22. What is the 60th percentile for the this data?

a. 2
b. 3
c. 4
d. 5

Use the following information to answer the next two exercises: The following data are the results of a random survey of 110 Reservists called to active duty to increase security at California airports.

Number of Dependents	Frequency
0	11
1	27
2	33
3	20
4	19

23. Construct a 95% confidence interval for the true population mean number of dependents of Reservists called to active duty to increase security at California airports.

a. (1.85, 2.32)
b. (1.80, 2.36)
c. (1.97, 2.46)
d. (1.92, 2.50)

24. The 95% confidence interval above means:

a. Five percent of confidence intervals constructed this way will not contain the true population aveage number of dependents.
b. We are 95% confident the true population mean number of dependents falls in the interval.
c. Both of the above answer choices are correct.
d. None of the above.

25. $X \sim U(4, 10)$. Find the 30$^{\text{th}}$ percentile.

a. 0.3000
b. 3
c. 5.8
d. 6.1

26. If $X \sim Exp(0.8)$, then $P(x < \mu)$ = __________

a. 0.3679
b. 0.4727
c. 0.6321
d. cannot be determined

27. The lifetime of a computer circuit board is normally distributed with a mean of 2,500 hours and a standard deviation of 60 hours. What is the probability that a randomly chosen board will last at most 2,560 hours?

a. 0.8413
b. 0.1587
c. 0.3461
d. 0.6539

28. A survey of 123 reservists called to active duty as a result of the September 11, 2001, attacks was conducted to determine the proportion that were married. Eighty-six reported being married. Construct a 98% confidence interval for the true population proportion of reservists called to active duty that are married.

a. (0.6030, 0.7954)
b. (0.6181, 0.7802)
c. (0.5927, 0.8057)
d. (0.6312, 0.7672)

29. Winning times in 26 mile marathons run by world class runners average 145 minutes with a standard deviation of 14 minutes. A sample of the last ten marathon winning times is collected. Let x = mean winning times for ten marathons. The distribution for x is:

a. $N\left(145, \frac{14}{\sqrt{10}}\right)$
b. $N\,(145{,}14)$
c. t_9
d. t_{10}

30. Suppose that Phi Beta Kappa honors the top one percent of college and university seniors. Assume that grade point means (GPA) at a certain college are normally distributed with a 2.5 mean and a standard deviation of 0.5. What would be the minimum GPA needed to become a member of Phi Beta Kappa at that college?

a. 3.99
b. 1.34
c. 3.00
d. 3.66

The number of people living on American farms has declined steadily during the 20$^{\text{th}}$ century. Here are data on the farm population (in millions of persons) from 1935 to 1980.

Year	1935	1940	1945	1950	1955	1960	1965	1970	1975	198
Population	32.1	30.5	24.4	23.0	19.1	15.6	12.4	9.7	8.9	7.2

31. The linear regression equation is $\hat{y}$ = 1166.93 – 0.5868*x*. What was the expected farm population (in millions of persons) for 1980?

a. 7.2
b. 5.1
c. 6.0
d. 8.0

32. In linear regression, which is the best possible *SSE*?

a. 13.46
b. 18.22
c. 24.05
d. 16.33

33. In regression analysis, if the correlation coefficient is close to one what can be said about the best fit line?

a. It is a horizontal line. Therefore, we can not use it.
b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.
c. The coefficient correlation is close to the limit. Therefore, it is hard to make a decision.
d. We do not have the equation. Therefore, we cannot say anything about it.

Use the following information to answer the next three exercises: A study of the career plans of young women and men sent questionnaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen. Here are the data from the students who responded.

	Female	**Male**
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Does the data suggest that there is a relationship between the gender of students and their choice of major?

34. The distribution for the test is:

a. $\text{Chi}^2{}_8$.
b. $\text{Chi}^2{}_3$.
c. t_{721}.
d. $N(0, 1)$.

35. The expected number of female who choose finance is:

a. 37.
b. 61.
c. 60.
d. 70.

36. The *p*-value is 0.0127 and the level of significance is 0.05. The conclusion to the test is:

a. there is insufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
b. there is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.
c. there is sufficient evidence to conclude that students find economics very hard.
d. there is in sufficient evidence to conclude that more females prefer administration than males.

37. An agency reported that the work force nationwide is composed of 10% professional, 10% clerical, 30% skilled, 15% service, and 35% semiskilled laborers. A random sample of 100 San Jose residents indicated 15 professional, 15 clerical, 40 skilled, 10 service, and 20 semiskilled laborers. At $\alpha = 0.10$ does the work force in San Jose appear to be consistent with the agency report for the nation? Which kind of test is it?

a. Chi^2 goodness of fit
b. Chi^2 test of independence
c. Independent groups proportions
d. Unable to determine

Practice Final Exam 1 Solutions

Solutions

1. b. independent

2. c. $\frac{4}{16}$

3. b. Two measurements are drawn from the same pair of individuals or objects.

4. b. $\frac{68}{118}$

5. d. $\frac{30}{52}$

6. b. $\frac{8}{40}$

7. b. 2.78

8. a. 8.25

9. c. 0.2870

10. c. Normal

11. d. $H_a: p_A \neq p_B$

12. b. conclude that the pass rate for Math 1A is different than the pass rate for Math 1B when, in fact, the pass rates are the same.

13. b. not reject H_0

14. c. Iris

15. c. Student's *t*

16. b. is left-tailed.

17. c. cluster sampling

18. b. median

19. a. the probability that an outcome of the data will happen purely by chance when the null hypothesis is true.

20. d. stratified

21. b. 25

22. c. 4

23. a. (1.85, 2.32)

24. c. Both above are correct.

25. c. 5.8

26. c. 0.6321

27. a. 0.8413

28. a. (0.6030, 0.7954)

29. a. $N\left(145, \frac{14}{\sqrt{10}}\right)$

30. d. 3.66

31. b. 5.1

32. a. 13.46

33. b. There is a strong linear pattern. Therefore, it is most likely a good model to be used.

34. b. $\text{Chi}^2{}_3$.

35. d. 70

36. b. There is sufficient evidence to conclude that the choice of major and the gender of the student are not independent of each other.

37. a. Chi^2 goodness-of-fit

Practice Final Exam 2

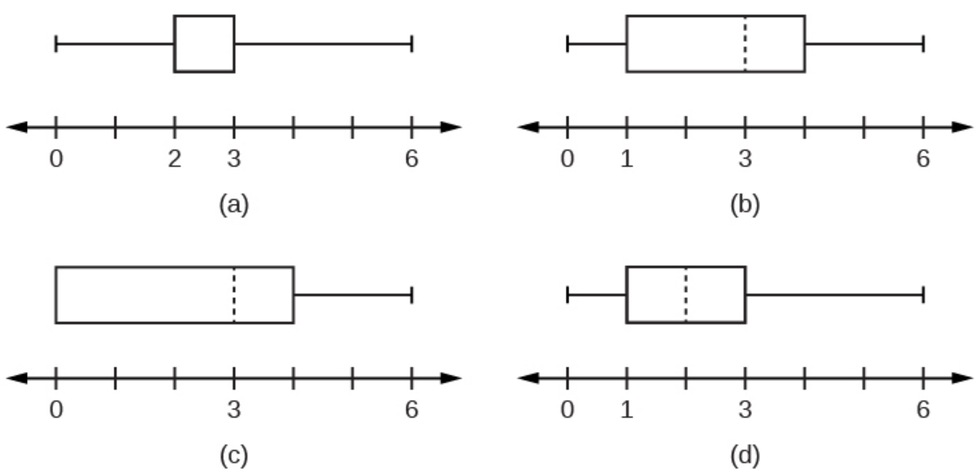
1. A study was done to determine the proportion of teenagers that own a car. The population proportion of teenagers that own a car is the:

a. statistic.
b. parameter.
c. population.
d. variable.

Use the following information to answer the next two exercises:

value	frequency
0	1
1	4
2	7
3	9
6	4

2. The box plot for the data is:



3. If six were added to each value of the data in the table, the 15th percentile of the new list of values is:

a. six
b. one
c. seven
d. eight

Use the following information to answer the next two exercises: Suppose that the probability of a drought in any independent year is 20%. Out of those years in which a drought occurs, the probability of water rationing is ten percent. However, in any year, the probability of water rationing is five percent.

4. What is the probability of both a drought and water rationing occurring?

a. 0.05
b. 0.01

c. 0.02
d. 0.30

5. Which of the following is true?

a. Drought and water rationing are independent events.
b. Drought and water rationing are mutually exclusive events.
c. None of the above

Use the following information to answer the next two exercises: Suppose that a survey yielded the following data:

gender	**apple**	**pumpkin**	**pecan**
female	40	10	30
male	20	30	10

Favorite Pie

6. Suppose that one individual is randomly chosen. The probability that the person's favorite pie is apple or the person is male is _____.

a. $\frac{40}{60}$
b. $\frac{60}{140}$
c. $\frac{120}{140}$
d. $\frac{100}{140}$

7. Suppose H_0 is: Favorite pie and gender are independent. The *p*-value is ______.

a. ≈ 0
b. 1
c. 0.05
d. cannot be determined

Use the following information to answer the next two exercises: Let's say that the probability that an adult watches the news at least once per week is 0.60. We randomly survey 14 people. Of interest is the number of people who watch the news at least once per week.

8. Which of the following statements is FALSE?

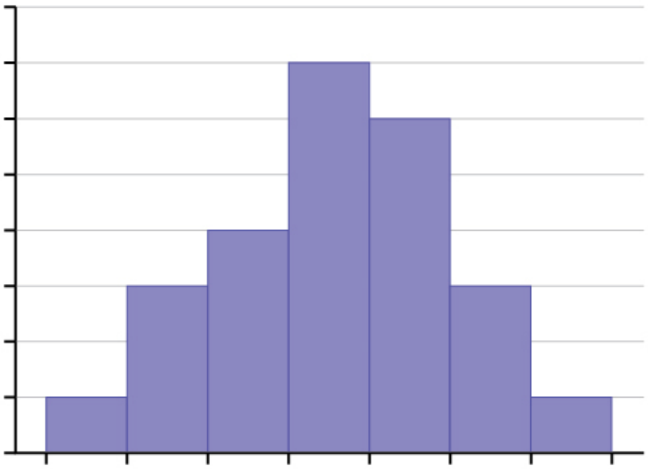
a. $X \sim B(14\ 0.60)$
b. The values for x are: {1 ,2 ,3 ,... ,14}.
c. $\mu = 8.4$
d. $P(X = 5) = 0.0408$

9. Find the probability that at least six adults watch the news at least once per week.

a. $\frac{6}{14}$
b. 0.8499
c. 0.9417

d. 0.6429

10. The following histogram is most likely to be a result of sampling from which distribution?



a. chi-square with *df* = 6
b. exponential
c. uniform
d. binomial

11. The ages of campus day and evening students is known to be normally distributed. A sample of six campus day and evening students reported their ages (in years) as: {18, 35, 27, 45, 20, 20}. What is the error bound for the 90% confidence interval of the true average age?

a. 11.2
b. 22.3
c. 17.5
d. 8.7

12. If a normally distributed random variable has $\mu = 0$ and $\sigma = 1$, then 97.5% of the population values lie above:

a. –1.96.
b. 1.96.
c. 1.
d. –1.

Use the following information to answer the next three exercises. The amount of money a customer spends in one trip to the supermarket is known to have an exponential distribution. Suppose the average amount of money a customer spends in one trip to the supermarket is $72.

13. What is the probability that one customer spends less than $72 in one trip to the supermarket?

a. 0.6321
b. 0.5000
c. 0.3714
d. 1

14. How much money altogether would you expect the next five customers to spend in one trip to the supermarket (in dollars)?

a. 72
b. $\frac{72^2}{5}$
c. 5184
d. 360

15. If you want to find the probability that the mean amount of money 50 customers spend in one trip to the supermarket is less than $60, the distribution to use is:

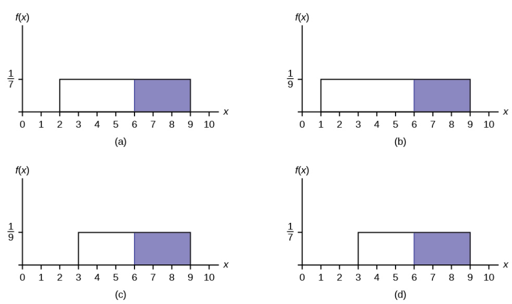
a. $N(72, 72)$
b. $N\left(72, \frac{72}{\sqrt{50}}\right)$
c. $Exp(72)$
d. $Exp\left(\frac{1}{72}\right)$

Use the following information to answer the next three exercises: The amount of time it takes a fourth grader to carry out the trash is uniformly distributed in the interval from one to ten minutes.

16. What is the probability that a randomly chosen fourth grader takes more than seven minutes to take out the trash?

a. $\frac{3}{9}$
b. $\frac{7}{9}$
c. $\frac{3}{10}$
d. $\frac{7}{10}$

17. Which graph best shows the probability that a randomly chosen fourth grader takes more than six minutes to take out the trash given that he or she has already taken more than three minutes?



18. We should expect a fourth grader to take how many minutes to take out the trash?

a. 4.5
b. 5.5
c. 5
d. 10

Use the following information to answer the next three exercises: At the beginning of the quarter, the amount of time a student waits in line at the campus cafeteria is normally distributed with a mean of five minutes and a standard deviation of 1.5 minutes.

19. What is the 90th percentile of waiting times (in minutes)?

a. 1.28

b. 90
c. 7.47
d. 6.92

20. The median waiting time (in minutes) for one student is:

a. 5.
b. 50.
c. 2.5.
d. 1.5.

21. Find the probability that the average wait time for ten students is at most 5.5 minutes.

a. 0.6301
b. 0.8541
c. 0.3694
d. 0.1459

22. A sample of 80 software engineers in Silicon Valley is taken and it is found that 20% of them earn approximately $50,000 per year. A point estimate for the true proportion of engineers in Silicon Valley who earn $50,000 per year is:

a. 16.
b. 0.2.
c. 1.
d. 0.95.

23. If $P(Z < z_\alpha) = 0.1587$ where $Z \sim N(0, 1)$, then α is equal to:

a. –1.
b. 0.1587.
c. 0.8413.
d. 1.

24. A professor tested 35 students to determine their entering skills. At the end of the term, after completing the course, the same test was administered to the same 35 students to study their improvement. This would be a test of:

a. independent groups.
b. two proportions.
c. matched pairs, dependent groups.
d. exclusive groups.

A math exam was given to all the third grade children attending ABC School. Two random samples of scores were taken.

	n	x	***s***
Boys	55	82	5
Girls	60	86	7

25. Which of the following correctly describes the results of a hypothesis test of the claim, “There is a difference between the mean scores obtained by third grade girls and boys at the 5% level of significance”?

a. Do not reject H_0. There is insufficient evidence to conclude that there is a difference in the mean scores.
b. Do not reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.
c. Reject H_0. There is insufficient evidence to conclude that there is no difference in the mean scores.
d. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. In a survey of 80 males, 45 had played an organized sport growing up. Of the 70 females surveyed, 25 had played an organized sport growing up. We are interested in whether the proportion for males is higher than the proportion for females. The correct conclusion is that:

a. there is insufficient information to conclude that the proportion for males is the same as the proportion for females.
b. there is insufficient information to conclude that the proportion for males is not the same as the proportion for females.
c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.
d. not enough information to make a conclusion.

27. From past experience, a statistics teacher has found that the average score on a midterm is 81 with a standard deviation of 5.2. This term, a class of 49 students had a standard deviation of 5 on the midterm. Do the data indicate that we should reject the teacher’s claim that the standard deviation is 5.2? Use $\alpha = 0.05$.

a. Yes
b. No
c. Not enough information given to solve the problem

28. Three loading machines are being compared. Ten samples were taken for each machine. Machine I took an average of 31 minutes to load packages with a standard deviation of two minutes. Machine II took an average of 28 minutes to load packages with a standard deviation of 1.5 minutes. Machine III took an average of 29 minutes to load packages with a standard deviation of one minute. Find the *p*-value when testing that the average loading times are the same.

a. *p*-value is close to zero
b. *p*-value is close to one
c. not enough information given to solve the problem

Use the following information to answer the next three exercises: A corporation has offices in different parts of the country. It has gathered the following information concerning the number of bathrooms and the number of employees at seven sites:

Number of employees *x*	650	730	810	900	102	107	1150
Number of bathrooms *y*	40	50	54	61	82	110	121

29. Is the correlation between the number of employees and the number of bathrooms significant?

a. Yes
b. No
c. Not enough information to answer question

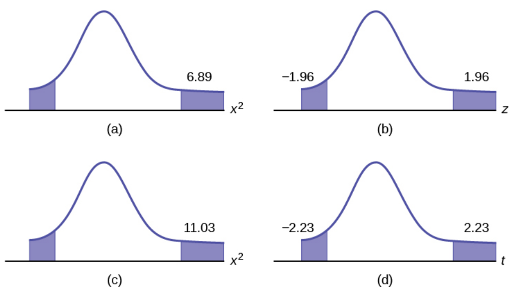
30. The linear regression equation is:

a. $\hat{y} = 0.0094 - 79.96x$
b. $\hat{y} = 79.96 + 0.0094x$
c. $\hat{y} = 79.96 - 0.0094x$
d. $\hat{y} = -0.0094 + 79.96x$

31. If a site has 1,150 employees, approximately how many bathrooms should it have?

a. 69
b. 91
c. 91,954
d. We should not be estimating here.

32. Suppose that a sample of size ten was collected, with $\bar{x} = 4.4$ and $s = 1.4$. H_0: $\sigma^2 = 1.6$ vs. H_a: $\sigma^2 \neq 1.6$. Which graph best describes the results of the test?



Sixty-four backpackers were asked the number of days since their latest backpacking trip. The number of days is given in [link]:

# of days	1	2	3	4	5	6	7	8
Frequency	5	9	6	12	7	10	5	10

33. Conduct an appropriate test to determine if the distribution is uniform.

a. The p-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.
b. The p-value is < 0.01. There is sufficient information to conclude the distribution is not uniform.
c. The p-value is between 0.01 and 0.10, but without alpha (α) there is not enough information
d. There is no such test that can be conducted.

34. Which of the following statements is true when using one-way ANOVA?

a. The populations from which the samples are selected have different distributions.
b. The sample sizes are large.
c. The test is to determine if the different groups have the same means.
d. There is a correlation between the factors of the experiment.

Practice Final Exam 2 Solutions

Solutions

1. b. parameter.

2. a.

3. c. seven

4. c. 0.02

5. c. none of the above

6. d. $\frac{100}{140}$

7. a. ≈ 0

8. b. The values for *x* are: {1, 2, 3,..., 14}

9. c. 0.9417.

10. d. binomial

11. d. 8.7

12. a. –1.96

13. a. 0.6321

14. d. 360

15. b. $N\left(72, \frac{72}{\sqrt{50}}\right)$

16. a. $\frac{3}{9}$

17. d.

18. b. 5.5

19. d. 6.92

20. a. 5

21. b. 0.8541

22. b. 0.2

23. a. –1.

24. c. matched pairs, dependent groups.

25. d. Reject H_0. There is sufficient evidence to conclude that there is a difference in the mean scores.

26. c. there is sufficient evidence to conclude that the proportion for males is higher than the proportion for females.

27. b. no

28. b. *p*-value is close to 1.

29. b. No

30. c. $\hat{y}$ = 79.96x – 0.0094

31. d. We should not be estimating here.

32. a.

33. a. The p-value is > 0.10. There is insufficient information to conclude that the distribution is not uniform.

34. c. The test is to determine if the different groups have the same means.

Data Sets

Lap Times

The following tables provide lap times from Terri Vogel's log book. Times are recorded in seconds for 2.5-mile laps completed in a series of races and practice runs.

	Lap 1	**Lap 2**	**Lap 3**	**Lap 4**	**Lap 5**	**Lap 6**	**Lap 7**
Race 1	135	130	131	132	130	131	133
Race 2	134	131	131	129	128	128	129
Race 3	129	128	127	127	130	127	129
Race 4	125	125	126	125	124	125	125
Race 5	133	132	132	132	131	130	132
Race 6	130	130	130	129	129	130	129
Race 7	132	131	133	131	134	134	131
Race 8	127	128	127	130	128	126	128

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Race 9	132	130	127	128	126	127	124
Race 10	135	131	131	132	130	131	130
Race 11	132	131	132	131	130	129	129
Race 12	134	130	130	130	131	130	130
Race 13	128	127	128	128	128	129	128
Race 14	132	131	131	131	132	130	130
Race 15	136	129	129	129	129	129	129
Race 16	129	129	129	128	128	129	129
Race 17	134	131	132	131	132	132	132
Race 18	129	129	130	130	133	133	127
Race 19	130	129	129	129	129	129	128
Race 20	131	128	130	128	129	130	130

Race Lap Times (in seconds)

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 1	142	143	180	137	134	134	172
Practice 2	140	135	134	133	128	128	131
Practice 3	130	133	130	128	135	133	133
Practice 4	141	136	137	136	136	136	145
Practice 5	140	138	136	137	135	134	134
Practice 6	142	142	139	138	129	129	127
Practice 7	139	137	135	135	137	134	135
Practice 8	143	136	134	133	134	133	132
Practice 9	135	134	133	133	132	132	133
Practice 10	131	130	128	129	127	128	127

	Lap 1	Lap 2	Lap 3	Lap 4	Lap 5	Lap 6	Lap 7
Practice 11	143	139	139	138	138	137	138
Practice 12	132	133	131	129	128	127	126
Practice 13	149	144	144	139	138	138	137
Practice 14	133	132	137	133	134	130	131
Practice 15	138	136	133	133	132	131	131

Practice Lap Times (in seconds)

Stock Prices

The following table lists initial public offering (IPO) stock prices for all 1999 stocks that at least doubled in value during the first day of trading.

$17.00	$23.00	$14.00	$16.00	$12.00	$26.00
$20.00	$22.00	$14.00	$15.00	$22.00	$18.00
$18.00	$21.00	$21.00	$19.00	$15.00	$21.00
$18.00	$17.00	$15.00	$25.00	$14.00	$30.00
$16.00	$10.00	$20.00	$12.00	$16.00	$17.44

$16.00	$14.00	$15.00	$20.00	$20.00	$16.00
$17.00	$16.00	$15.00	$15.00	$19.00	$48.00
$16.00	$18.00	$9.00	$18.00	$18.00	$20.00
$8.00	$20.00	$17.00	$14.00	$11.00	$16.00
$19.00	$15.00	$21.00	$12.00	$8.00	$16.00
$13.00	$14.00	$15.00	$14.00	$13.41	$28.00
$21.00	$17.00	$28.00	$17.00	$19.00	$16.00
$17.00	$19.00	$18.00	$17.00	$15.00	
$14.00	$21.00	$12.00	$18.00	$24.00	
$15.00	$23.00	$14.00	$16.00	$12.00	
$24.00	$20.00	$14.00	$14.00	$15.00	
$14.00	$19.00	$16.00	$38.00	$20.00	
$24.00	$16.00	$8.00	$18.00	$17.00	
$16.00	$15.00	$7.00	$19.00	$12.00	
$8.00	$23.00	$12.00	$18.00	$20.00	
$21.00	$34.00	$16.00	$26.00	$14.00	

IPO Offer Prices

References

Data compiled by Jay R. Ritter of University of Florida using data from *Securities Data Co.* and *Bloomberg*.

Group and Partner Projects

Univariate Data

Student Learning Objectives

- The student will design and carry out a survey.
- The student will analyze and graphically display the results of the survey.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.
_____ Decide what data you are going to study.

Note:
Here are two examples, but you may **NOT** use them: number of M&M's per bag, number of pencils students have in their backpacks.

_____ Are your data discrete or continuous? How do you know?
_____ Decide how you are going to collect the data (for instance, buy 30 bags of M&M's; collect data from the World Wide Web).
_____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. Which method did you use? Why did you pick that method?
_____ Conduct your survey. **Your data size must be at least 30.**
_____ Summarize your data in a chart with columns showing **data value,**

frequency, relative frequency and cumulative relative frequency.
Answer the following (rounded to two decimal places):

a. = _____
b. s = _____
c. First quartile = _____
d. Median = _____
e. 70th percentile = _____

____ What value is two standard deviations above the mean?

____ What value is 1.5 standard deviations below the mean?
____ Construct a histogram displaying your data.
____ In complete sentences, describe the shape of your graph.
____ Do you notice any potential outliers? If so, what values are they? Show your work in how you used the potential outlier formula to determine whether or not the values might be outliers.
____ Construct a box plot displaying your data.
____ Does the middle 50% of the data appear to be concentrated together or spread apart? Explain how you determined this.
____ Looking at both the histogram and the box plot, discuss the distribution of your data.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:

- **____Cover sheet**: name, class time, and name of your study
- **____Summary page**: This should contain paragraphs written with complete sentences. It should include answers to all the questions above. It should also include statements describing the population under study, the sample, a parameter or parameters being studied, and the statistic or statistics produced.
- **____URL** for data, if your data are from the World Wide Web

- **____Chart of data, frequency, relative frequency, and cumulative relative frequency**
- **____Page(s) of graphs:** histogram and box plot

Continuous Distributions and Central Limit Theorem

Student Learning Objectives

- The student will collect a sample of continuous data.
- The student will attempt to fit the data sample to various distribution models.
- The student will validate the central limit theorem.

Instructions

As you complete each task below, check it off. Answer all questions in your summary.

Part I: Sampling

____ Decide what **continuous** data you are going to study. (Here are two examples, but you may NOT use them: the amount of money a student spent on college supplies this term, or the length of time distance telephone call lasts.)
____ Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random (using a random number generator) sampling. Do not use convenience sampling. What method did you use? Why did you pick that method?
____ Conduct your survey. Gather **at least 150 pieces of continuous, quantitative data**.
____ Define (in words) the random variable for your data. *X* = ________
____ Create two lists of your data: (1) unordered data, (2) in order of smallest to largest.

_____ Find the sample mean and the sample standard deviation (rounded to two decimal places).

a. = _______
b. s = _______

_____ Construct a histogram of your data containing five to ten intervals of equal width. The histogram should be a representative display of your data. Label and scale it.

Part II: Possible Distributions

_____ Suppose that X followed the following theoretical distributions. Set up each distribution using the appropriate information from your data.
_____ Uniform: $X \sim U$ ______________ Use the lowest and highest values as a and b.
_____ Normal: $X \sim N$ ______________ Use to estimate for μ and s to estimate for σ.
_____ **Must** your data fit one of the above distributions? Explain why or why not.
_____ **Could** the data fit two or three of the previous distributions (at the same time)? Explain.
_____ Calculate the value k(an X value) that is 1.75 standard deviations above the sample mean. k = __________ (rounded to two decimal places)
Note: k = + (1.75)s
_____ Determine the relative frequencies (RF) rounded to four decimal places.

Note:
Note
____

a. $RF(X < k)$ = _______
b. $RF(X > k)$ = _______
c. $RF(X = k)$ = _______

Note:
Note
You should have one page for the uniform distribution, one page for the exponential distribution, and one page for the normal distribution.

_____ State the distribution: $X \sim$ __________
_____ Draw a graph for each of the three theoretical distributions. Label the axes and mark them appropriately.
_____ Find the following theoretical probabilities (rounded to four decimal places).

a. $P(X < k)$ = _______
b. $P(X > k)$ = _______
c. $P(X = k)$ = _______

_____ Compare the relative frequencies to the corresponding probabilities. Are the values close?
_____ Does it appear that the data fit the distribution well? Justify your answer by comparing the probabilities to the relative frequencies, and the histograms to the theoretical graphs.

Part III: CLT Experiments

_______ From your original data (before ordering), use a random number generator to pick 40 samples of size five. For each sample, calculate the average.
_______ On a separate page, attached to the summary, include the 40 samples of size five, along with the 40 sample averages.

_______ List the 40 averages in order from smallest to largest.
_______ Define the random variable, , in words. = _________________
_______ State the approximate theoretical distribution of . ~
____
_______ Base this on the mean and standard deviation from your original data.
_______ Construct a histogram displaying your data. Use five to six intervals of equal width. Label and scale it.
Calculate the value (an value) that is 1.75 standard deviations above the sample mean. = ______ (rounded to two decimal places)
Determine the relative frequencies (*RF*) rounded to four decimal places.

a. *RF*(<) = ________
b. *RF*(>) = ________
c. *RF*(=) = ________

Find the following theoretical probabilities (rounded to four decimal places).

a. *P*(<) = ________
b. *P*(>) = ________
c. *P*(=) = ________

_______ Draw the graph of the theoretical distribution of .
_______ Compare the relative frequencies to the probabilities. Are the values close?
_______ Does it appear that the data of averages fit the distribution of well? Justify your answer by comparing the probabilities to the relative frequencies, and the histogram to the theoretical graph.
In three to five complete sentences for each, answer the following questions. Give thoughtful explanations.
_______ In summary, do your original data seem to fit the uniform, exponential, or normal distributions? Answer why or why not for each distribution. If the data do not fit any of those distributions, explain why.
_______ What happened to the shape and distribution when you averaged your data? **In theory,** what should have happened? In theory, would “it”

always happen? Why or why not?
______ Were the relative frequencies compared to the theoretical probabilities closer when comparing the or distributions? Explain your answer.

Assignment Checklist

You need to turn in the following typed and stapled packet, with pages in the following order:
____ **Cover sheet**: name, class time, and name of your study
____ **Summary pages**: These should contain several paragraphs written with complete sentences that describe the experiment, including what you studied and your sampling technique, as well as answers to all of the questions previously asked questions
____ **URL** for data, if your data are from the World Wide Web
____ **Pages, one for each theoretical distribution**, with the distribution stated, the graph, and the probability questions answered
____ **Pages of the data requested**
____ **All graphs required**

Hypothesis Testing-Article

Student Learning Objectives

- The student will identify a hypothesis testing problem in print.
- The student will conduct a survey to verify or dispute the results of the hypothesis test.
- The student will summarize the article, analysis, and conclusions in a report.

Instructions

As you complete each task, check it off. Answer all questions in your summary.
____**Find an article** in a newspaper, magazine, or on the internet which makes a claim about **ONE** population mean or **ONE** population proportion. The claim may be based upon a survey that the article was reporting on. Decide whether this claim is the null or alternate hypothesis.
____**Copy or print out the article** and include a copy in your project, along with the source.
____**State how you will collect your data.** (Convenience sampling is not acceptable.)
____**Conduct your survey. You must have more than 50 responses in your sample.** When you hand in your final project, attach the tally sheet or the packet of questionnaires that you used to collect data. Your data must be real.
____**State the statistics** that are a result of your data collection: sample size, sample mean, and sample standard deviation, OR sample size and number of successes.
____**Make two copies of the appropriate solution sheet.**
____**Record the hypothesis test** on the solution sheet, based on your experiment. **Do a DRAFT solution** first on one of the solution sheets and check it over carefully. Have a classmate check your solution to see if it is done correctly. Make your decision using a 5% level of significance. Include the 95% confidence interval on the solution sheet.
____**Create a graph that illustrates your data.** This may be a pie or bar graph or may be a histogram or box plot, depending on the nature of your data. Produce a graph that makes sense for your data and gives useful visual information about your data. You may need to look at several types of graphs before you decide which is the most appropriate for the type of data in your project.
____**Write your summary** (in complete sentences and paragraphs, with proper grammar and correct spelling) that describes the project. The summary **MUST** include:

a. Brief discussion of the article, including the source
b. Statement of the claim made in the article (one of the hypotheses).
c. Detailed description of how, where, and when you collected the data, including the sampling technique; did you use cluster, stratified,

systematic, or simple random sampling (using a random number generator)? As previously mentioned, convenience sampling is not acceptable.

d. Conclusion about the article claim in light of your hypothesis test; this is the conclusion of your hypothesis test, stated in words, in the context of the situation in your project in sentence form, as if you were writing this conclusion for a non-statistician.
e. Sentence interpreting your confidence interval in the context of the situation in your project

Assignment Checklist

Turn in the following typed (12 point) and stapled packet for your final project:

____**Cover sheet** containing your name(s), class time, and the name of your study

____**Summary**, which includes all items listed on summary checklist

____**Solution sheet** neatly and completely filled out. The solution sheet does not need to be typed.

____**Graphic representation of your data**, created following the guidelines previously discussed; include only graphs which are appropriate and useful.

____**Raw data collected AND a table summarizing the sample data** (n, and s; or x, n, and p', as appropriate for your hypotheses); the raw data does not need to be typed, but the summary does. Hand in the data as you collected it. (Either attach your tally sheet or an envelope containing your questionnaires.)

Bivariate Data, Linear Regression, and Univariate Data

Student Learning Objectives

- The students will collect a bivariate data sample through the use of appropriate sampling techniques.
- The student will attempt to fit the data to a linear model.

- The student will determine the appropriateness of linear fit of the model.
- The student will analyze and graph univariate data.

Instructions

1. As you complete each task below, check it off. Answer all questions in your introduction or summary.
2. Check your course calendar for intermediate and final due dates.
3. Graphs may be constructed by hand or by computer, unless your instructor informs you otherwise. All graphs must be neat and accurate.
4. All other responses must be done on the computer.
5. Neatness and quality of explanations are used to determine your final grade.

Part I: Bivariate Data

Introduction
____State the bivariate data your group is going to study.

> **Note:**
> Here are two examples, but you may **NOT** use them: height vs. weight and age vs. running distance.

____Describe your sampling technique in detail. Use cluster, stratified, systematic, or simple random sampling (using a random number generator) sampling. Convenience sampling is **NOT** acceptable.
____Conduct your survey. Your number of pairs must be at least 30.
____Print out a copy of your data.

Analysis
____On a separate sheet of paper construct a scatter plot of the data. Label and scale both axes.
____State the least squares line and the correlation coefficient.
____On your scatter plot, in a different color, construct the least squares line.
____Is the correlation coefficient significant? Explain and show how you determined this.
____Interpret the slope of the linear regression line in the context of the data in your project. Relate the explanation to your data, and quantify what the slope tells you.
____Does the regression line seem to fit the data? Why or why not? If the data does not seem to be linear, explain if any other model seems to fit the data better.
____Are there any outliers? If so, what are they? Show your work in how you used the potential outlier formula in the Linear Regression and Correlation chapter (since you have bivariate data) to determine whether or not any pairs might be outliers.

Part II: Univariate Data

In this section, you will use the data for **ONE** variable only. Pick the variable that is more interesting to analyze. For example: if your independent variable is sequential data such as year with 30 years and one piece of data per year, your *x*-values might be 1971, 1972, 1973, 1974, …, 2000. This would not be interesting to analyze. In that case, choose to use the dependent variable to analyze for this part of the project.
______Summarize your data in a chart with columns showing data value, frequency, relative frequency, and cumulative relative frequency.
______Answer the following question, rounded to two decimal places:

a. Sample mean = _______
b. Sample standard deviation = _______
c. First quartile = _______
d. Third quartile = _______
e. Median = _______

f. 70th percentile = _______
g. Value that is 2 standard deviations above the mean = _______
h. Value that is 1.5 standard deviations below the mean = _______

______Construct a histogram displaying your data. Group your data into six to ten intervals of equal width. Pick regularly spaced intervals that make sense in relation to your data. For example, do NOT group data by age as 20-26,27-33,34-40,41-47,48-54,55-61 . . . Instead, maybe use age groups 19.5-24.5, 24.5-29.5, . . . or 19.5-29.5, 29.5-39.5, 39.5-49.5, . . .
______In complete sentences, describe the shape of your histogram.
______Are there any potential outliers? Which values are they? Show your work and calculations as to how you used the potential outlier formula in Descriptive Statistics (since you are now using univariate data) to determine which values might be outliers.
______Construct a box plot of your data.
______Does the middle 50% of your data appear to be concentrated together or spread out? Explain how you determined this.
______Looking at both the histogram AND the box plot, discuss the distribution of your data. For example: how does the spread of the middle 50% of your data compare to the spread of the rest of the data represented in the box plot; how does this correspond to your description of the shape of the histogram; how does the graphical display show any outliers you may have found; does the histogram show any gaps in the data that are not visible in the box plot; are there any interesting features of your data that you should point out.

Due Dates

- Part I, Intro: ___________ (keep a copy for your records)
- Part I, Analysis: ___________ (keep a copy for your records)

- Entire Project, typed and stapled: ___________

_____ Cover sheet: names, class time, and name of your study

_____ Part I: label the sections “Intro” and “Analysis.”

____ Part II:

____ Summary page containing several paragraphs written in complete sentences describing the experiment, including what you studied and how you collected your data. The summary page should also include answers to ALL the questions asked above.

____ All graphs requested in the project

____ All calculations requested to support questions in data

____ Description: what you learned by doing this project, what challenges you had, how you overcame the challenges

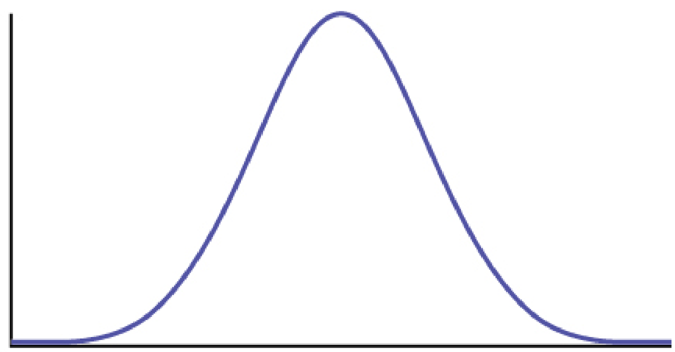
Note:
Note
Include answers to ALL questions asked, even if not explicitly repeated in the items above.

Solution Sheets

Hypothesis Testing with One Sample

Class Time: ___________________________
Name: _______________________________________

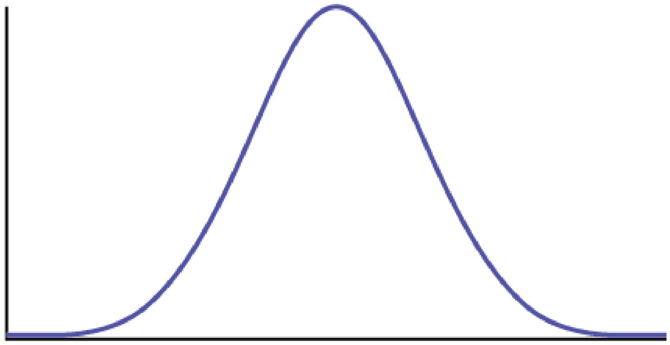
a. H_0: ________
b. H_a: ________
c. In words, **CLEARLY** state what your random variable or represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one or two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. CLEARLY, label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

   i. Alpha: ________
   ii. Decision: ________
   iii. Reason for decision: ________
   iv. Conclusion: ________

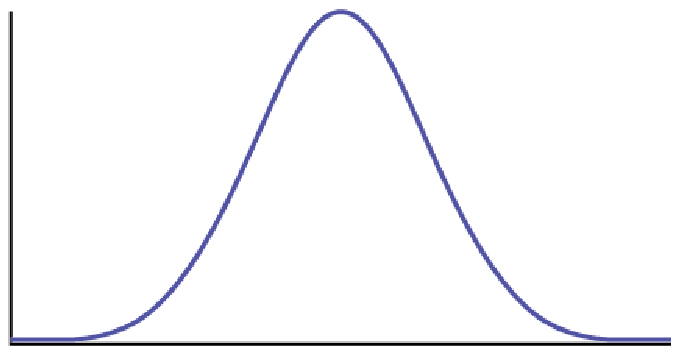
i. Construct a 95% confidence interval for the true mean or proportion. Include a sketch of the graph of the situation. Label the point estimate and the lower and upper bounds of the confidence interval.



Hypothesis Testing with Two Samples

Class Time: ___________________________
Name: ______________________________________

a. H_0: ________
b. H_a: ________
c. In words, **clearly** state what your random variable ,
or represents.
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the p-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **CLEARLY** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis), the reason for it, and write an appropriate conclusion, using **complete sentences**.

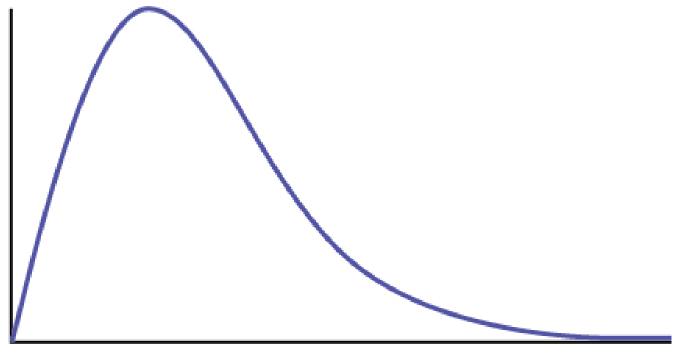
   a. Alpha: ________
   b. Decision: ________
   c. Reason for decision: ________
   d. Conclusion: ________

i. In complete sentences, explain how you determined which distribution to use.

The Chi-Square Distribution

Class Time: ____________________________
Name: ____

a. H_0: ________
b. H_a: ________
c. What are the degrees of freedom?
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value? In one to two complete sentences, explain what the *p*-value means for this problem.
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.



h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences.**

   i. Alpha: ________
   ii. Decision: ________
   iii. Reason for decision: ________
   iv. Conclusion: ________

F Distribution and One-Way ANOVA

Class Time: ____________________________

Name: _______________________________________

a. H_0: ________
b. H_a: ________
c. $df(n)$ = _______ $df(d)$ = ________
d. State the distribution to use for the test.
e. What is the test statistic?
f. What is the *p*-value?
g. Use the previous information to sketch a picture of this situation. **Clearly** label and scale the horizontal axis and shade the region(s) corresponding to the *p*-value.

h. Indicate the correct decision ("reject" or "do not reject" the null hypothesis) and write appropriate conclusions, using **complete sentences**.

a. Alpha: ________
b. Decision: ________
c. Reason for decision: ________
d. Conclusion: ________

Mathematical Phrases, Symbols, and Formulas

English Phrases Written Mathematically

When the English says:	**Interpret this as:**
X is at least 4.	$X \geq 4$
The minimum of X is 4.	$X \geq 4$
X is no less than 4.	$X \geq 4$
X is greater than or equal to 4.	$X \geq 4$
X is at most 4.	$X \leq 4$
The maximum of X is 4.	$X \leq 4$
X is no more than 4.	$X \leq 4$
X is less than or equal to 4.	$X \leq 4$
X does not exceed 4.	$X \leq 4$
X is greater than 4.	$X > 4$
X is more than 4.	$X > 4$
X exceeds 4.	$X > 4$
X is less than 4.	$X < 4$

When the English says:	Interpret this as:
There are fewer *X* than 4.	$X < 4$
X is 4.	$X = 4$
X is equal to 4.	$X = 4$
X is the same as 4.	$X = 4$
X is not 4.	$X \neq 4$
X is not equal to 4.	$X \neq 4$
X is not the same as 4.	$X \neq 4$
X is different than 4.	$X \neq 4$

Formulas

Formula 1: Factorial

$n! = n(n-1)(n-2)\dots(1)$

$0! = 1$

Formula 2: Combinations

$\binom{n}{r} = \frac{n!}{(n-r)!r!}$

Formula 3: Binomial Distribution

$X \sim B(n, p)$

$P(X = x) = \binom{n}{x} p^x q^{n-x}$, for $x = 0, 1, 2, \ldots, n$

Formula 4: Geometric Distribution

$X \sim G(p)$

$P(X = x) = q^{x-1} p$, for $x = 1, 2, 3, \ldots$

Formula 5: Hypergeometric Distribution

$X \sim H(r, b, n)$

$P(X = x) = \left(\frac{\binom{r}{x}\binom{b}{n-x}}{\binom{r+b}{n}} \right)$

Formula 6: Poisson Distribution

$X \sim P(\mu)$

$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$

Formula 7: Uniform Distribution

$X \sim U(a, b)$

$f(X) = \frac{1}{b-a}, a < x < b$

Formula 8: Exponential Distribution

$X \sim Exp(m)$

$f(x) = me^{-mx} m > 0, x \geq 0$

Formula 9: Normal Distribution
$X \sim N(\mu, \sigma^2)$

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$

Formula 10: Gamma Function

$$\Gamma(z) = \int_{\infty}^{0} x^{z-1} e^{-x} dx \ z > 0$$

$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$\Gamma(m+1) = m!$ for m, a nonnegative integer

otherwise: $\Gamma(a+1) = a\Gamma(a)$

Formula 11: Student's *t*-distribution

$X \sim t_{df}$

$$f(x) = \frac{\left(1+\frac{x^2}{n}\right)^{\frac{-(n+1)}{2}} \Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\mathrm{n\pi}}\,\Gamma\left(\frac{n}{2}\right)}$$

$$X = \frac{Z}{\sqrt{\frac{Y}{n}}}$$

$Z \sim N(0,1\ ),\ Y \sim X^2_{df}$, n = degrees of freedom

Formula 12: Chi-Square Distribution

$X \sim X^2_{df}$

$f(x) = \frac{x^{\frac{n-2}{2}} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)}$, $x > 0$, n = positive integer and degrees of freedom

Formula 13: F Distribution

$X \sim F_{df(n),df(d)}$

$df(n) =$degrees of freedom for the numerator

$df(d) =$degrees of freedom for the denominator

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)}\left(\frac{u}{v}\right)^{\frac{u}{2}} x^{\left(\frac{u}{2}-1\right)}\left[1+\left(\frac{u}{v}\right)x^{-0.5(u+v)}\right]$$

$X = \frac{Y_u}{W_v}$, Y, W are chi-square

Symbols and Their Meanings

Chapter (1st used)	Symbol	Spoken	Meaning
Sampling and Data	$\sqrt{}$	The square root of	same
Sampling and Data	π	Pi	3.14159… (a specific number)
Descriptive Statistics	Q_1	Quartile one	the first quartile
Descriptive Statistics	Q_2	Quartile two	the second quartile
Descriptive Statistics	Q_3	Quartile three	the third quartile
Descriptive Statistics	IQR	interquartile range	$Q_3 - Q_1 = IQR$
Descriptive Statistics	$\overline{x}$	x-bar	sample mean
Descriptive Statistics	μ	mu	population mean

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Descriptive Statistics	**s** s_x sx	s	sample standard deviation
Descriptive Statistics	s^2 s_x^2	s squared	sample variance
Descriptive Statistics	σ σ_x σx	sigma	population standard deviation
Descriptive Statistics	σ^2 σ_x^2	sigma squared	population variance
Descriptive Statistics	Σ	capital sigma	sum
Probability Topics	$\{\}$	brackets	set notation
Probability Topics	S	S	sample space
Probability Topics	A	Event A	event A
Probability Topics	$P(A)$	probability of A	probability of A occurring

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Probability Topics	$P(A\|B)$	probability of A given B	prob. of A occurring given B has occurred
Probability Topics	$P(A \text{ OR } B)$	prob. of A or B	prob. of A or B or both occurring
Probability Topics	$P(A \text{ AND } B)$	prob. of A and B	prob. of both A and B occurring (same time)
Probability Topics	A'	A-prime, complement of A	complement of A, not A
Probability Topics	$P(A')$	prob. of complement of A	same
Probability Topics	G_1	green on first pick	same
Probability Topics	$P(G_1)$	prob. of green on first pick	same
Discrete Random Variables	PDF	prob. distribution function	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	X	X	the random variable X
Discrete Random Variables	$X \sim$	the distribution of X	same
Discrete Random Variables	B	binomial distribution	same
Discrete Random Variables	G	geometric distribution	same
Discrete Random Variables	H	hypergeometric dist.	same
Discrete Random Variables	P	Poisson dist.	same
Discrete Random Variables	λ	Lambda	average of Poisson distribution
Discrete Random Variables	$\geq$	greater than or equal to	same

Chapter (1st used)	Symbol	Spoken	Meaning
Discrete Random Variables	$\leq$	less than or equal to	same
Discrete Random Variables	$=$	equal to	same
Discrete Random Variables	$\neq$	not equal to	same
Continuous Random Variables	$f(x)$	f of x	function of x
Continuous Random Variables	*pdf*	prob. density function	same
Continuous Random Variables	U	uniform distribution	same
Continuous Random Variables	*Exp*	exponential distribution	same
Continuous Random Variables	k	k	critical value

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Continuous Random Variables	$f(x) =$	f of x equals	same
Continuous Random Variables	m	m	decay rate (for exp. dist.)
The Normal Distribution	N	normal distribution	same
The Normal Distribution	z	z-score	same
The Normal Distribution	Z	standard normal dist.	same
The Central Limit Theorem	CLT	Central Limit Theorem	same
The Central Limit Theorem	$\overline{X}$	X-bar	the random variable X-bar
The Central Limit Theorem	μ_x	mean of X	the average of X

Chapter (1st used)	Symbol	Spoken	Meaning
The Central Limit Theorem	$\mu_{\overline{x}}$	mean of *X*-bar	the average of *X*-bar
The Central Limit Theorem	σ_x	standard deviation of *X*	same
The Central Limit Theorem	$\sigma_{\overline{x}}$	standard deviation of *X*-bar	same
The Central Limit Theorem	ΣX	sum of *X*	same
The Central Limit Theorem	Σx	sum of *x*	same
Confidence Intervals	*CL*	confidence level	same
Confidence Intervals	*CI*	confidence interval	same
Confidence Intervals	*EBM*	error bound for a mean	same
Confidence Intervals	*EBP*	error bound for a proportion	same

Chapter (1st used)	Symbol	Spoken	Meaning
Confidence Intervals	t	Student's *t*-distribution	same
Confidence Intervals	df	degrees of freedom	same
Confidence Intervals	$t_{\frac{\alpha}{2}}$	student t with *a*/2 area in right tail	same
Confidence Intervals	$p\prime$; $\widehat{p}$	*p*-prime; *p*-hat	sample proportion of success
Confidence Intervals	$q\prime$; $\hat{q}$	*q*-prime; *q*-hat	sample proportion of failure
Hypothesis Testing	H_0	*H*-naught, *H*-sub 0	null hypothesis
Hypothesis Testing	H_a	*H*-*a*, *H*-sub *a*	alternate hypothesis
Hypothesis Testing	H_1	*H*-1, *H*-sub 1	alternate hypothesis
Hypothesis Testing	α	alpha	probability of Type I error

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Hypothesis Testing	β	beta	probability of Type II error
Hypothesis Testing	$\overline{X1} - \overline{X2}$	*X*1-bar minus *X*2-bar	difference in sample means
Hypothesis Testing	$\mu_1 - \mu_2$	*mu*-1 minus *mu*-2	difference in population means
Hypothesis Testing	$P'_1 - P'_2$	*P*1-prime minus *P*2-prime	difference in sample proportions
Hypothesis Testing	$p_1 - p_2$	*p*1 minus *p*2	difference in population proportions
Chi-Square Distribution	X^2	*Ky*-square	Chi-square
Chi-Square Distribution	O	Observed	Observed frequency
Chi-Square Distribution	E	Expected	Expected frequency

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
Linear Regression and Correlation	$y = a + bx$	y equals a plus b-x	equation of a line
Linear Regression and Correlation	$\hat{y}$	y-hat	estimated value of y
Linear Regression and Correlation	r	correlation coefficient	same
Linear Regression and Correlation	ε	error	same
Linear Regression and Correlation	SSE	Sum of Squared Errors	same
Linear Regression and Correlation	$1.9s$	1.9 times s	cut-off value for outliers

Chapter (1st used)	**Symbol**	**Spoken**	**Meaning**
F-Distribution and ANOVA	F	F-ratio	F-ratio

Symbols and their Meanings

Notes for the TI-83, 83+, 84, 84+ Calculators

Quick Tips

Legend

- 

  represents a button press
- [ ] represents yellow command or green letter behind a key
- < > represents items on the screen

To adjust the contrast
Press

, then hold

to increase the contrast or

to decrease the contrast.

To capitalize letters and words
Press

to get one capital letter, or press

, then

to set all button presses to capital letters. You can return to the top-level button values by pressing

ALPHA

again.

To correct a mistake
If you hit a wrong button, just hit

CLEAR

and start again.

To write in scientific notation
Numbers in scientific notation are expressed on the TI-83, 83+, 84, and 84+ using E notation, such that...

- 4.321 E 4 = 4.321×10^4
- 4.321 E –4 = 4.321×10^{-4}

To transfer programs or equations from one calculator to another:
Both calculators: Insert your respective end of the link cable cable and press

2nd

, then `[LINK]`.

Calculator receiving information:

Use the arrows to navigate to and select `<RECEIVE>`
Press ENTER.

Calculator sending information:

Press appropriate number or letter.
Use up and down arrows to access the appropriate item.

Press ENTER to select item to transfer.

Press right arrow to navigate to and select `<TRANSMIT>`.
Press ENTER.

> **Note:**
> Note
> ERROR 35 LINK generally means that the cables have not been inserted far enough.

Both calculators: Insert your respective end of the link cable cable Both calculators: press

2nd

, then `[QUIT]` to exit when done.

Manipulating One-Variable Statistics

> **Note:**
> Note
> These directions are for entering data with the built-in statistical program.

Data	Frequency

Data	Frequency
–2	10
–1	3
0	4
1	5
3	8

Sample DataWe are manipulating one-variable statistics.

To begin:

1. Turn on the calculator.

2. Access statistics mode.

3. Select `<4:ClrList>` to clear data from lists, if desired.

   ,

4. Enter list `[L1]` to be cleared.

   , `[L1]` ,

5. Display last instruction.

   2nd

   , [ENTRY]

6. Continue clearing remaining lists in the same fashion, if desired.

   ◀

   ,

   2nd

   , [L2] ,

   ENTER

7. Access statistics mode.

   STAT

8. Select <1:Edit . . .>

   ENTER

9. Enter data. Data values go into [L1]. (You may need to arrow over to [L1]).

   - Type in a data value and enter it. (For negative numbers, use the negate (-) key at the bottom of the keypad).

     (−)

     ,

     9

     ,

     ENTER

- Continue in the same manner until all data values are entered.

10. In `[L2]`, enter the frequencies for each data value in `[L1]`.

    - Type in a frequency and enter it. (If a data value appears only once, the frequency is "1").

      4

      ,

      ENTER

    - Continue in the same manner until all data values are entered.

11. Access statistics mode.

    STAT

12. Navigate to `<CALC>`.

13. Access `<1:1-var Stats>`.

    ENTER

14. Indicate that the data is in `[L1]`...

    2nd

    , `[L1]` ,

    ,

15. ...and indicate that the frequencies are in `[L2]`.

    2nd

    , `[L2]` ,

    ENTER

16. The statistics should be displayed. You may arrow down to get remaining statistics. Repeat as necessary.

Drawing Histograms

> **Note:**
> Note
> We will assume that the data is already entered.

We will construct two histograms with the built-in STATPLOT application. The first way will use the default ZOOM. The second way will involve customizing a new graph.

1. Access graphing mode.

   , `[STAT PLOT]`

2. Select `<1:plot 1>` to access plotting - first graph.

3. Use the arrows navigate go to `<ON>` to turn on Plot 1.
   `<ON>` ,

4. Use the arrows to go to the histogram picture and select the histogram.

5. Use the arrows to navigate to `<Xlist>`.

6. If "L1" is not selected, select it.

2nd

, [L1] ,

ENTER

7. Use the arrows to navigate to <Freq>.

8. Assign the frequencies to [L2].

2nd

, [L2] ,

ENTER

9. Go back to access other graphs.

2nd

, [STAT PLOT]

10. Use the arrows to turn off the remaining plots.
11. **Be sure to deselect or clear all equations before graphing.**

To deselect equations:

1. Access the list of equations.

Y=

2. Select each equal sign (=).

▼

▶

ENTER

3. Continue, until all equations are deselected.

To clear equations:

1. Access the list of equations.

2. Use the arrow keys to navigate to the right of each equal sign (=) and clear them.

3. Repeat until all equations are deleted.

To draw default histogram:

1. Access the ZOOM menu.

2. Select `<9:ZoomStat>`.

3. The histogram will show with a window automatically set.

To draw custom histogram:

1. Access window mode to set the graph parameters.

2.
   - $X_{\text{min}} = -2.5$
   - $X_{\text{max}} = 3.5$
   - $X_{scl} = 1$ (width of bars)
   - $Y_{\text{min}} = 0$
   - $Y_{\text{max}} = 10$
   - $Y_{scl} = 1$ (spacing of tick marks on *y*-axis)
   - $X_{res} = 1$

3. Access graphing mode to see the histogram.

   GRAPH

To draw box plots:

1. Access graphing mode.

   2nd

   , `[STAT PLOT]`

2. Select `<1:Plot 1>` to access the first graph.

   ENTER

3. Use the arrows to select `<ON>` and turn on Plot 1.

   ENTER

4. Use the arrows to select the box plot picture and enable it.

   ENTER

5. Use the arrows to navigate to `<Xlist>`.

6. If "L1" is not selected, select it.

   2nd

   , `[L1]` ,

   ENTER

7. Use the arrows to navigate to `<Freq>`.

8. Indicate that the frequencies are in `[L2]`.

   2nd

   , `[L2]` ,

ENTER

9. Go back to access other graphs.

   2nd

   , `[STAT PLOT]`
10. **Be sure to deselect or clear all equations before graphing** using the method mentioned above.
11. View the box plot.

    GRAPH

    , `[STAT PLOT]`

Linear Regression

Sample Data

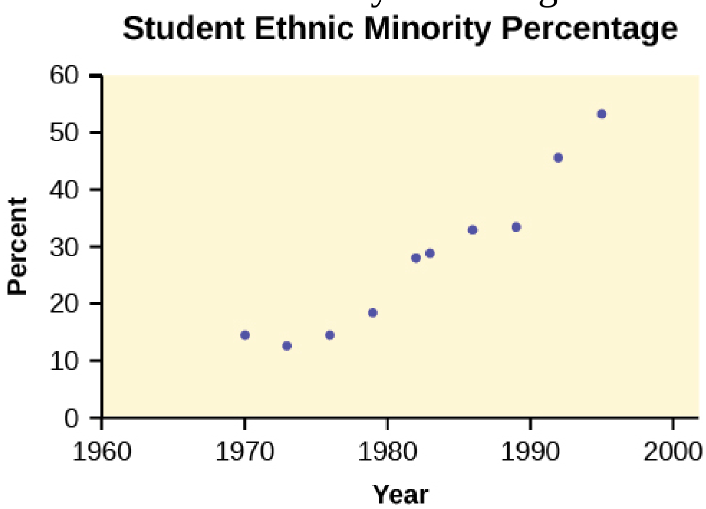
The following data is real. The percent of declared ethnic minority students at De Anza College for selected years from 1970–1995 was:

Year	**Student Ethnic Minority Percentage**
1970	14.13
1973	12.27
1976	14.08
1979	18.16

Year	Student Ethnic Minority Percentage
1982	27.64
1983	28.72
1986	31.86
1989	33.14
1992	45.37
1995	53.1

The independent variable is "Year," while the independent variable is "Student Ethnic Minority Percent."

Student Ethnic Minority Percentage



By hand, verify the scatterplot above.

Note:

Note
The TI-83 has a built-in linear regression feature, which allows the data to be edited.The *x*-values will be in `[L1]`; the *y*-values in `[L2]`.

To enter data and do linear regression:

1. ON Turns calculator on.

2. Before accessing this program, be sure to turn off all plots.

   - Access graphing mode.

     , `[STAT PLOT]`

   - Turn off all plots.

     ,

3. Round to three decimal places. To do so:

   - Access the mode menu.

     , `[STAT PLOT]`

   - Navigate to `<Float>` and then to the right to `<3>`.

    - All numbers will be rounded to three decimal places until changed.

      ENTER

4. Enter statistics mode and clear lists `[L1]` and `[L2]`, as describe previously.

   STAT

   ,

   4

5. Enter editing mode to insert values for *x* and *y*.

   STAT

   ,

   ENTER

6. Enter each value. Press

   ENTER

   to continue.

To display the correlation coefficient:

1. Access the catalog.

   2nd

   , `[CATALOG]`

2. Arrow down and select `<DiagnosticOn>`

   ▼

   ... ,

ENTER

,

ENTER

3. r and r^2 will be displayed during regression calculations.

4. Access linear regression.

STAT

▶

5. Select the form of *y* = *a* + *bx*.

8

,

ENTER

The display will show:

LinReg

- *y* = *a* + *bx*
- *a* = –3176.909
- *b* = 1.617
- *r* = 2 0.924
- *r* = 0.961

This means the Line of Best Fit (Least Squares Line) is:

- *y* = –3176.909 + 1.617*x*
- Percent = –3176.909 + 1.617 (year #)

The correlation coefficient *r* = 0.961

To see the scatter plot:

1. Access graphing mode.

   2nd

   , [STAT PLOT]

2. Select <1:plot 1> To access plotting - first graph.

   ENTER

3. Navigate and select <ON> to turn on Plot 1.
   <ON>

   ENTER

4. Navigate to the first picture.

5. Select the scatter plot.

   ENTER

6. Navigate to <Xlist>.
7. If [L1] is not selected, press

   2nd

   , [L1] to select it.

8. Confirm that the data values are in [L1].
   <ON>

   ENTER

9. Navigate to <Ylist>.

10. Select that the frequencies are in [L2].

    2nd

, [L2] ,

ENTER

11. Go back to access other graphs.

2nd

, [STAT PLOT]

12. Use the arrows to turn off the remaining plots.
13. Access window mode to set the graph parameters.

WINDOW

- $X_{\min} = 1970$
- $X_{\max} = 2000$
- $X_{scl} = 10$ (spacing of tick marks on *x*-axis)
- $Y_{\min} = -0.05$
- $Y_{\max} = 60$
- $Y_{scl} = 10$ (spacing of tick marks on *y*-axis)
- $X_{res} = 1$

14. Be sure to deselect or clear all equations before graphing, using the instructions above.
15. Press the graph button to see the scatter plot.

GRAPH

To see the regression graph:

1. Access the equation menu. The regression equation will be put into Y1.

Y=

2. Access the vars menu and navigate to <5: Statistics>.

VARS

,

5

3. Navigate to <EQ>.

4. <1: RegEQ> contains the regression equation which will be entered in Y1.

   ENTER

5. Press the graphing mode button. The regression line will be superimposed over the scatter plot.

   GRAPH

To see the residuals and use them to calculate the critical point for an outlier:

1. Access the list. RESID will be an item on the menu. Navigate to it.

   2nd

   , [LIST], <RESID>

2. Confirm twice to view the list of residuals. Use the arrows to select them.

   ENTER

   ,

   ENTER

3. The critical point for an outlier is: $1.9V\frac{\text{SSE}}{n-2}$ where:

   - n = number of pairs of data
   - SSE = sum of the squared errors
   - $\sum \text{residual}^2$

4. Store the residuals in [L3].

   STO►

,

2nd

, `[L3]` ,

ENTER

5. Calculate the $\frac{(\text{residual})^2}{n-2}$. Note that $n-2=8$

2nd

, `[L3]` ,

x^2

,

÷

,

8

6. Store this value in `[L4]`.

STO►

,

2nd

, `[L4]` ,

ENTER

7. Calculate the critical value using the equation above.

1

,

.

,

9

,

×

,

2nd

, [V] ,

2nd

, [LIST]

▶

,

▶

,

5

,

2nd

, [L4] ,

)

,

)

,

ENTER

8. Verify that the calculator displays: 7.642669563. This is the critical value.
9. Compare the absolute value of each residual value in `[L3]` to 7.64. If the absolute value is greater than 7.64, then the (x, y) corresponding point is an outlier. In this case, none of the points is an outlier.

To obtain estimates of *y* for various *x*-values:
There are various ways to determine estimates for "*y*." One way is to substitute values for "*x*" in the equation. Another way is to use the

TRACE

on the graph of the regression line.

TI-83, 83+, 84, 84+ instructions for distributions and tests

Distributions

Access `DISTR` (for "Distributions").

For technical assistance, visit the Texas Instruments website at http://www.ti.com and enter your calculator model into the "search" box.

Binomial Distribution

- `binompdf(n,p,x)` corresponds to $P(X = x)$
- `binomcdf(n,p,x)` corresponds to $P(X \le x)$
- To see a list of all probabilities for *x*: 0, 1, . . . , *n*, leave off the "`x`" parameter.

Poisson Distribution

- `poissonpdf(λ,x)` corresponds to $P(X = x)$
- `poissoncdf(λ,x)` corresponds to $P(X \le x)$

Continuous Distributions (general)

- $-\infty$ uses the value –1EE99 for left bound
- ∞ uses the value 1EE99 for right bound

Normal Distribution

- `normalpdf(x,μ,σ)` yields a probability density function value (only useful to plot the normal curve, in which case "`x`" is the variable)
- `normalcdf(left bound, right bound, μ, σ)` corresponds to P(left bound < X < right bound)
- `normalcdf(left bound, right bound)` corresponds to P(left bound < Z < right bound) – standard normal
- `invNorm(p,μ,σ)` yields the critical value, k: $P(X < k) = p$
- `invNorm(p)` yields the critical value, k: $P(Z < k) = p$ for the standard normal

Student's *t*-Distribution

- `tpdf(x,df)` yields the probability density function value (only useful to plot the student-*t* curve, in which case "`x`" is the variable)
- `tcdf(left bound, right bound, df)` corresponds to P(left bound < t < right bound)

Chi-square Distribution

- `X²pdf(x,df)` yields the probability density function value (only useful to plot the chi^2 curve, in which case "`x`" is the variable)
- `X²cdf(left bound, right bound, df)` corresponds to P(left bound < X^2 < right bound)

F Distribution

- `Fpdf(x,dfnum,dfdenom)` yields the probability density function value (only useful to plot the F curve, in which case "`x`" is the variable)
- `Fcdf(left bound,right bound,dfnum,dfdenom)` corresponds to P(left bound < F < right bound)

Tests and Confidence Intervals

Access `STAT` and `TESTS`.

For the confidence intervals and hypothesis tests, you may enter the data into the appropriate lists and press `DATA` to have the calculator find the sample means and standard deviations. Or, you may enter the sample means and sample standard deviations directly by pressing `STAT` once in the appropriate tests.

Confidence Intervals

- `ZInterval` is the confidence interval for mean when σ is known.
- `TInterval` is the confidence interval for mean when σ is unknown; *s* estimates σ.
- `1-PropZInt` is the confidence interval for proportion.

> **Note:**
> Note
> The confidence levels should be given as percents (ex. enter "`95`" or "`.95`" for a 95% confidence level).

Hypothesis Tests

- `Z-Test` is the hypothesis test for single mean when σ is known.
- `T-Test` is the hypothesis test for single mean when σ is unknown; *s* estimates σ.
- `2-SampZTest` is the hypothesis test for two independent means when both σ's are known.
- `2-SampTTest` is the hypothesis test for two independent means when both σ's are unknown.
- `1-PropZTest` is the hypothesis test for single proportion.
- `2-PropZTest` is the hypothesis test for two proportions.
- X^2`-Test` is the hypothesis test for independence.

- `X²GOF-Test` is the hypothesis test for goodness-of-fit (TI-84+ only).
- `LinRegTTEST` is the hypothesis test for Linear Regression (TI-84+ only).

Note:
Note
Input the null hypothesis value in the row below "`Inpt`." For a test of a single mean, "`μ∅`" represents the null hypothesis. For a test of a single proportion, "`p∅`" represents the null hypothesis. Enter the alternate hypothesis on the bottom row.

Tables

The module contains links to government site tables used in statistics.

Note:
Note
When you are finished with the table link, use the back button on your browser to return here.

Tables (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 3, 2009)

- Student *t* table
- Normal table
- Chi-Square table
- F-table
- All four tables can be accessed by going to

95% Critical Values of the Sample Correlation Coefficient Table

- 95% Critical Values of the Sample Correlation Coefficient